nuclei specifically direct the synthesis by Pol III of a series of discrete, homologous RNAs of 100–110 nucleotides which probably correspond to BC2. These could be early terminating species or perhaps precursors of BC1 which are not polyadenylated in this system. Although there is no precedence for polyadenylation of Pol III transcripts, our evidence suggests that BC1 contains poly(A), and that BC2 does not and is a Pol III product. Post-transcriptional polyadenylation could account for the difference in size of the 110-nucleotide α-amanitin-resistant S100-extract product and the 160-nucleotide BC1 species found *in vivo*. One model for the generation of BC1 and BC2 RNA is that ID sequences with their 3′ oligo(dA) tracts are located in brain genes. These are recognized by neural transcription specificity factors and then transcribed by Pol III, with termination occurring past the oligo(dA) region to generate BC2. Most such BC2 transcripts are unstable. A few of the 10^5 ID sequences, however, may have a T in the oligo(dA) region, forming the sequence AATAAA which can serve as a poly(A) addition signal, thereby producing BC1. The polyadenylated BC1 molecules are transported to the cytoplasm where they form a stable pool of a few thousand molecules which may or may not be functional. Meanwhile, we speculate that the Pol III transcription of ID sequences has activated the genes containing the ID sequence in a necessary manner for Pol II transcription and consequently specific mRNA expression.

Clearly, ID sequences could not be the only factor involved in controlling neural gene expression because the brain is a collection of heterogeneous cell types which are not homogeneous in terms of mRNA expression[9,20]. ID sequences may be necessary but are not sufficient for brain gene expression. They perhaps represent a general primary neural gene signal upon which other control systems operate. For most brain-specific genes, a secondary positive signal is required for expression. Specificity is also mediated by salt-releasable factors which keep transcription by Pol II of regions containing ID sequences at low levels in non-brain tissues, arguing that negative control, acting either at the level of specific repressors or, more likely, at the level of overall chromatin structure, is responsible for keeping neural-specific genetic regions silent in non-neural tissues. ID sequences may be the lineage-specific control elements.

This model, then, utilizes *cis*-acting sequences within introns to identify tissue-specific genes, and *trans*-acting transcription specificity factors to guide Pol III to those genes. A separate switching mechanism must regulate the tissue-specific expression of the transcription specificity factors. Liver and kidney apparently produce transcripts analogous to BC2 but having different sequences. These may be transcripts of the liver and kidney *cis*-acting elements.

Clearly our prejudice in this discussion has been that ID sequences influence transcription; however, they could alternatively or additionally be involved in mRNA processing as nucleation sites for RNA folding or protein–RNA interactions (as discussed by Milner *et al.*[4]). The specificity of BC1 RNA suggests it, too, could carry out a brain-specific role.

We thank Judy Ogata, Chris Clare, Mary Ann Brow and Debbie Andrews for technical contributions. This work was supported in part by grants to J.G.S. from NIH (GM 32355) and McNeil Laboratories, and to J.M.G. from NIH (GM 26453). This is publication no. 2988-IMM of the Research Institute of Scripps Clinic.

Received 30 September; accepted 12 December 1983.

1. Sutcliffe, J. G., Milner, R. J., Bloom, F. E. & Lerner, R. A. *Proc. natn. Acad. Sci. U.S.A.* **79,** 4942–4946 (1982).
2. Barta, A., Richards, R. I., Baxter, J. D. & Shine, J. *Proc. natn. Acad. Sci. U.S.A.* **78,** 4867–4871 (1981).
3. Hojvat, S., Baker, G., Kirsteins, L. & Lawrence, A. M. *Brain Res.* **239,** 543–557 (1982).
4. Milner, R. J., Bloom, F. E., Lai, C., Lerner, R. A. & Sutcliffe, J. G. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
5. Thomas, P. *Proc. natn. Acad. Sci. U.S.A.* **77,** 5201–5205 (1980).
6. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Bery, P. J. *Molec. Biol.* **113,** 237–251 (1977).
7. Reddy, R. *et al. J. biol. Chem.* **256,** 8452–8457 (1981).
8. Jelinek, W. & Leinwand, L. *Cell* **15,** 205–214 (1978).
9. Milner, R. J. & Sutcliffe, J. G. *Nucleic Acids Res.* **11,** 5497–5520 (1983).
10. Galli, G., Hofstetter, H. & Birnstiel, M. L. *Nature* **294,** 626–631 (1981).
11. Weil, P. A., Luse, D. S., Segall, J. & Roeder, R. G. *Cell* **18,** 469–484 (1979).
12. Brown, D. D. & Gurdon, J. B. *Proc. natn. Acad. Sci. U.S.A.* **79,** 2064–2068 (1977).
13. Sutcliffe, J. G. *Cold Spring Harb. Symp. quant. Biol.* **43,** 77–90 (1978).
14. Bogenhagen, D. F. & Brown, D. D. *Cell* **24,** 261–270 (1981).
15. Manley, J. L., Fine, A., Cano, A., Sharp, P. A. & Gefter, M. L. *Proc. natn. Acad. Sci. U.S.A.* **77,** 3855–3859 (1980).
16. Sudgen, B. & Keller, W. *J. biol. Chem.* **248,** 3777–3788 (1973).
17. Spadafora, C., Ondet, P. & Chambon, P. *Eur. J. Biochem.* **100,** 225–235 (1979).
18. Weisbrod, S. & Weintraub, H. *Proc. natn. Acad. Sci. U.S.A.* **76,** 631–635 (1979).
19. Manley, J. L. & Colozzo, M. T. *Nature* **300,** 376–379 (1982).
20. Sutcliffe, J. G., Milner, R. J., Shinnick, T. M. & Bloom, F. E. *Cell* **33,** 671–682 (1983).
21. Levy, D. E., Lerner, R. A. & Wilson, M. C. *Proc. natn. Acad. Sci. U.S.A.* **79,** 5823–5827 (1982).

Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin *in vivo*

Marilyn Kozak

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA

To determine whether sequence context influences the ability of an AUG triplet to be recognized as an initiator codon by eukaryotic ribosomes, single nucleotide changes were introduced near the translational start site in a cloned preproinsulin gene. Maximum synthesis of preproinsulin occurred when a purine, preferably adenosine, was located three nucleotides upstream from the initiator codon. Adenosine is found most frequently in that position in natural mRNAs.

TRANSLATION begins at the 5′-proximal AUG triplet in about 90% of the eukaryotic mRNAs that have been sequenced. That observation, together with other circumstantial evidence reviewed previously[1], prompted the hypothesis that the 40S subunit of eukaryotic ribosomes might find the AUG initiator codon by 'scanning' the 5′ end of messenger RNA. A partial vindication of the scanning model has come from experiments carried out in cell-free extracts. By studying the binding of specially designed templates to wheat germ ribosomes, we have demonstrated that mRNA must have a free 5′ terminus in order to be functional[2], that 40S ribosomal subunits can be trapped upstream from the AUG initiator codon[3] and that 40S subunits can migrate[3,4]. Consistent with a scanning mechanism, the translation of certain viral mRNAs was arrested following hybridization with DNA fragments complementary to a portion of the 5′-noncoding region[5–7]. Recent experiments carried out *in vivo* further support the idea of scanning: ribosomes appear to initiate exclusively at the 5′-proximal AUG triplet in mRNAs that contain tandemly reiterated copies of the preproinsulin ribosome binding site[8]. Although it seems clear that the position of

an AUG triplet relative to the 5′ end of the mRNA has a major role in identifying it as a functional initiator codon, the notion of scanning is not incompatible with the idea that sequence context modulates the efficiency with which an AUG triplet is recognized as an initiator codon. Indeed, the distribution of nucleotides flanking functional initiator codons in eukaryotic mRNAs is decidedly nonrandom. From a comparison of the 5′-noncoding sequences of over 200 cellular[9] and some 50 viral mRNAs[10], $CC^{A}_{G}CCAUGG$ has recently been tentatively proposed as a consensus sequence for eukaryotic initiation sites. The most highly conserved features revealed by those surveys were adenosine (A) in position −3, (that is, three nucleotides upstream from the AUG triplet, which is numbered +1 to +3) and guanosine (G) in position +4. The significance of the conserved nucleotides flanking the AUG triplet has been tested to a limited extent by measuring the binding of synthetic oligonucleotides to ribosomes *in vitro*[10]. To assess further whether the translational machinery in eukaryotes is sensitive to the sequence context around a potential AUG initiator codon, I have introduced single nucleotide changes near the translational start site in a cloned preproinsulin gene. The yield of proinsulin was then monitored during transient expression of the mutant plasmids in monkey (COS-1) cells. Sequence changes around the initiator codon drastically affected the yield of proinsulin, as shown in the experiments described below.

Plasmid construction

Plasmid 255 (p255) is a shuttle vector that encodes and expresses the rat preproinsulin II gene. (The plasmid was donated by Dr Peter Lomedico.) Its structure has been described previously[11] and is briefly outlined in Fig. 1*a*. Initiation of transcription at the simian virus 40 (SV40) early promoter in p255 gives rise to a chimaeric mRNA with a 5′-untranslated sequence of about 128 nucleotides: the first ~80 nucleotides are encoded by SV40 DNA and the remainder are from the rat insulin gene. A *Hin*dIII site marks the boundary between SV40 and rat insulin DNA, as shown in Fig. 1*a*, *b*. Rather than attempt to introduce mutations around the natural preproinsulin initiator codon, which lacks nearby restriction sites, I inserted upstream from the insulin coding sequence an 11-base pair (bp) oligonucleotide that carries another ATG triplet. The resulting mutant (p255/2) is illustrated in the second line of Fig. 1*c*, with the 11-nucleotide insert shown in bold face. To facilitate the construction and identification of additional mutants, the synthetic oligonucleotide was designed with *Hin*dIII and *Bcl*I restriction sites flanking the ATG triplet. Because the upstream ATG triplet in p255/2 lies in the same reading frame as the preproinsulin coding sequence, I anticipated that, after introducing mutations around the upstream ATG codon, I would be able to monitor initiation at the first versus the second ATG by measuring the size of the insulin-related polypeptides: the polypeptide resulting from initiation at the upstream ATG should carry an extra 18 amino acids at the N-terminus. That experimental design proved impractical, however, because the long form of preproinsulin encoded by p255/2 was still an efficient substrate for signal peptidase (M.K., unpublished data). Thus, irrespective of which ATG was used for initiation, the only polypeptide that accumulated was the cleavage product, proinsulin.

To obviate that problem, another derivative (p255A/1) was constructed, as illustrated in the bottom line of Fig. 1*c*. p255/2 DNA was cut with *Bcl*I and *Bam*HI which generate identical cohesive ends (GATC), although each enzyme recognizes a different hexameric sequence. The large DNA fragment was then re-ligated, that is, the *Bcl*I-generated end was joined to the *Bam*HI-generated end, thereby deleting 182 bp including the ATG triplet that is normally used to initiate translation of preproinsulin. In place of the natural initiator codon, p255A/1 has the ATG triplet derived from the oligonucleotide insert in p255/2. The sequence around the new initiator codon in p255A/1 is shown in Fig. 1*c*. p255A/1 directs synthesis of a polypeptide which lacks only three amino acids that are normally present at the N-terminus of preproinsulin. In the remainder of this article, p255A/1 is used as the point of reference: a series of mutants differing from p255A/1 by single nucleotide changes is described below, and their translational properties are compared.

Introduction of mutations

Site-specific mutagenesis was carried out by first forming gapped heteroduplex circles in which the region around the initiator codon was single-stranded (see Fig. 2*a*, *b*), and then annealing to the gap a complementary oligonucleotide that contains a single mismatched residue. The mismatched nucleotides are starred in Fig. 2. Two mutants were obtained using scheme *a*: p255/15 and p255/16 are identical to p255/2 (described above) except that p255/15 has an A and p255/16 has a G, in the −3 position. The two new mutants could not be used directly in the protein synthesis assay because the polypeptide resulting from initiation at the upstream ATG triplet would undergo cleavage, a problem explained in the preceding discussion of p255/2. Therefore, the technique whereby p255A/1 was derived from p255/2 was applied again: p255/15 was cut with *Bcl*I and *Bam*HI and then re-ligated, generating p255A/3; p255/16 was cut similarly and re-ligated to produce p255A/4. Derivatives p255A/3 and p255A/4 are identical in structure to p255A/1, except for the nucleotide in the −3 position. The sequence around the initiator codon in each plasmid is given in Table 1.

A slightly different scheme was needed to obtain mutants in the +4 position, because changing the A in that position would destroy the *Bcl*I site and thus preclude the *Bcl*I/*Bam*HI fusion step used to derive the aforementioned mutants. Derivative p255A/2, in which the sequence ATG$\underline{A}$ has been changed to ATG$\underline{G}$, was obtained using scheme *b* in Fig. 2. The technique again involved using a mismatched oligonucleotide to complement the single-stranded region in a gapped heteroduplex circle, but the parental plasmids used to construct the heteroduplex in scheme *b* differ from those in scheme *a*. Finally, scheme *c* in Fig. 2 illustrates the approach used to obtain a double mutant (p255A/7) in which the nucleotides in both the −3 and +4 positions have been changed to G. The technique here simply involved forming a nicked heteroduplex circle between two of the mutants already in hand, p255A/4 and p255A/2. The mismatched nucleotide in the −3 position of the heteroduplex underwent correction *in vivo* by the bacterial DNA repair machinery, generating the desired double mutant.

All of the mutants resulting from the constructions outlined in Fig. 2 were initially identified by the loss of the *Hin*dIII restriction site that had been present in the parental plasmids, or by the gain of an *Ava*II restriction site ($GG^{A}_{T}CC$) immediately following the ATG codon. The precise structure of each mutant was subsequently confirmed by DNA sequence analysis. The relevant portions of the sequencing gels are shown in Fig. 3.

Expression of mutant plasmids

Synthesis of proinsulin was monitored following transfection of COS-1 cells[12] by p255A/1 or one of the derivatives described above. Transfection of subconfluent cell monolayers was carried out using the standard $CaPO_4$–DNA co-precipitation technique of Wigler *et al.*[13], as described previously[8]. Two days after transfection the cells were labelled with ^{35}S-cysteine, lysed, and the insulin-related polypeptides recovered by precipitating with anti-insulin antibodies, all according to published procedures[8]. The product that accumulates when a cloned preproinsulin gene is introduced into cultured monkey cells has been reported by other workers to be proinsulin[14–16]. In the present study, identification of the proinsulin band rests on four criteria: it is absent from cells transfected with a control plasmid, p255/0, which lacks the entire insulin coding sequence; it is precipitated by anti-insulin antibodies but not by preimmune serum (not shown); it is competed by including excess nonradioactive insulin in the immunoprecipitation reaction (Fig. 4, lane 8); and its apparent molecular weight of ~9,000 corresponds approximately to that of proinsulin.

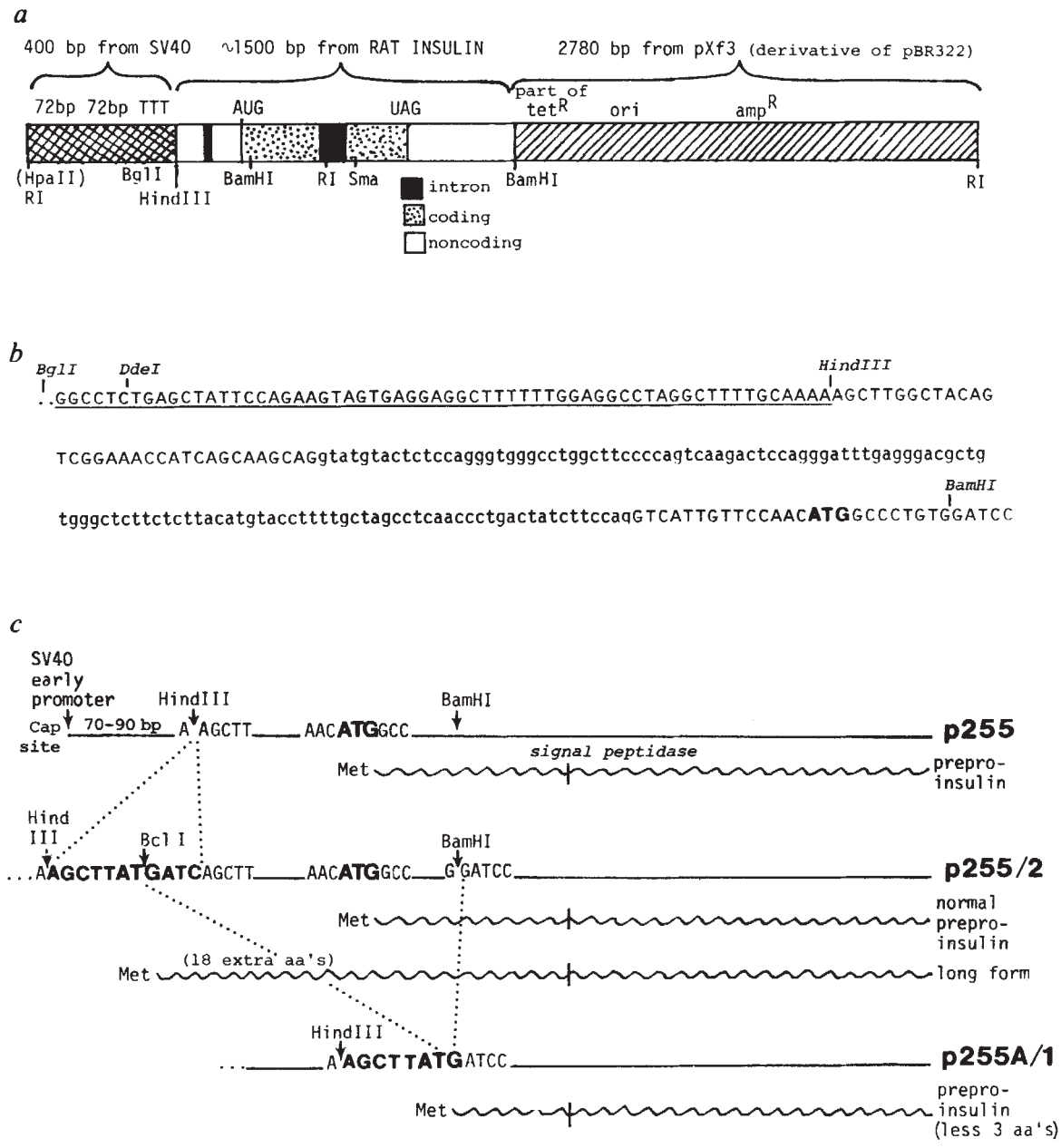

Fig. 1 Structure of transcripts and insulin-related polypeptides encoded by p255 and its derivatives. ***a***, Linear representation of the parental plasmid p255. The segment of SV40 DNA that is included in the plasmid extends from the *Hpa*II site at position 346 to the *Hin*dIII site at position 5,171, 8 bp upstream from the ATG initiator codon of T antigen. The plasmid contains a single *Hin*dIII site, which is at the junction of SV40 and rat insulin sequences. Some of the constructions described in the text were facilitated by eliminating the *Bam*HI site at the junction between rat genomic and pBR322 sequences in p255. This was done by obtaining unit-length linear DNA molecules after partial digestion with *Bam*HI, filling in the sticky ends using *Escherichia coli* DNA polymerase, and then re-ligating. The resulting plasmid (designated p255/11) has a single *Bam*HI site, 8 bp downstream from the ATG initiator codon of preproinsulin. In all other respects p255/11 is identical to wild-type p255. ***b***, Nucleotide sequence preceding and extending into the insulin coding region in p255. The DNA strand that has the same sense as mRNA is shown. The transcript produced by p255 initiates near the *Bgl*I site of SV40 and carries at its 5′ end 70–90 nucleotides encoded by SV40 DNA (underlined). The remainder of the transcript is derived from the rat insulin gene. The portion of the rat insulin sequence that is retained in mature mRNA is printed in capital letters without underlining; the 119-bp intron that interrupts the 5′-noncoding region is shown in lower case letters. The ATG codon that normally initiates translation of rat preproinsulin is shown in bold face. ***c***, Genesis of plasmids p255/2 and p255A/1. The upper line represents a portion of the parental plasmid p255. The *Hin*dIII and *Bam*HI sites are those identified in *b*. The second entry depicts p255/2, which was derived by inserting an 11-bp oligonucleotide (shown in bold face) at the unique *Hin*dIII site in p255. The third entry shows the structure of p255A/1, in which the sequence between the *Bcl*I and *Bam*HI sites has been deleted. The polypeptides encoded by each plasmid are represented by wavy lines.

Methods: The oligonucleotides 5′-AGCTTATGATC-3′ and 5′-AGCTGATCATA-3′ were custom-synthesized by PL Biochemicals, Inc. When annealed, they form a partially base-paired structure with 5′-protruding termini that are complementary to the overhangs generated by cutting DNA with *Hin*dIII. Derivative p255/2 was obtained by digesting the parental plasmid p255 with *Hin*dIII and alkaline phosphatase, followed by ligation with the aforementioned oligonucleotides which were present in approximately 50-fold molar excess. Plasmids that had acquired the 11-bp insert were detected initially by screening for the *Bcl*I cleavage site (TGATCA). Derivatives with the insert oriented correctly (that is, as illustrated in the figure) were subsequently identified by DNA sequence analysis. The deletion derivative p255A/1 was obtained by sequentially digesting p255/2 DNA (grown in the *dam*$^-$ host *E. coli* SR19) with *Bcl*I and *Bam*HI, and then re-ligating the linear DNA molecules. Plasmids from which the 182-bp *Bcl*I/*Bam*HI fragment had been deleted were initially identified by loss of the *Bcl*I and *Bam*HI restriction sites. The structure shown for p255A/1 was subsequently confirmed by DNA sequence analysis.

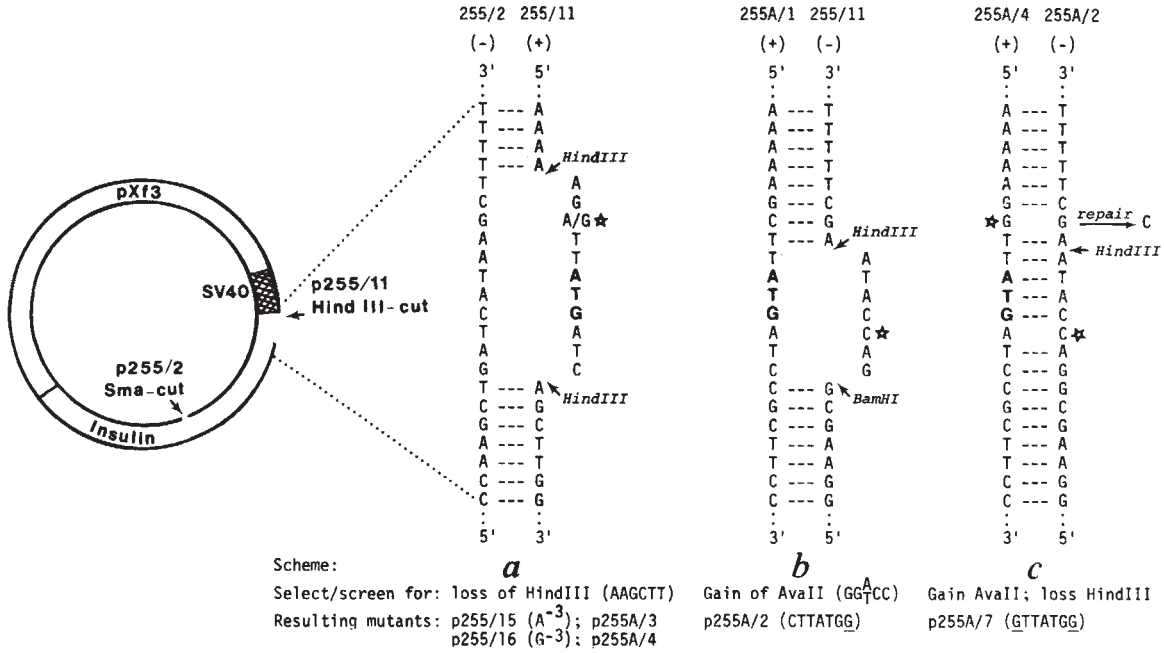

Fig. 2 Construction of point mutants. In protocols *a* and *b*, gapped heteroduplex circles were formed between two plasmid DNA strands of slightly dissimilar lengths, analogous to the procedure described in refs 19–21. The gapped circles were then annealed with an oligonucleotide that was complementary to the single-stranded region except for a single mismatched nucleotide, which is starred. In scheme *a*, for example, *Hin*dIII-cut p255/11 (which is identical to wild-type p255 except for loss of the *Bam*HI site at the junction between rat genomic and pXf3 DNA) and *Sma*I-cut p255/2 DNA were mixed, denatured and then annealed. The resulting heteroduplex circles, illustrated at the far left of the figure, have a single-stranded gap at the position of the 11-bp insertion in p255/2. In schemes *b* and *c*, the DNA strand shown on the left side of the heteroduplexes was linearized with *Sma*I, as in scheme *a*, but different nucleases (identified in the figure) were used to linearize the strand on the right. Recovery of the parental plasmids p255A/1 and p255A/4 was minimized in schemes *b* and *c* by using DNA from plasmids grown in the *dam*$^-$ host, *E. coli* SR19 (ref. 22). As shown in scheme *c*, the double mutant p255A/7 was obtained by forming a nicked heteroduplex circle between two of the mutants already in hand. *E. coli* cells were transformed with the nicked circles and repair occurred *in vivo* at the site of the mismatched nucleotide in position −3, just beyond the *Hin*dIII cut in the DNA strand derived from p255A/2. Each of the illustrated schemes gave only a low yield of the desired mutant; most of the ampicillin-resistant colonies were derived from one or the other unmodified parental plasmid. A portion of the parental-type background is attributable to the formation of gapped heteroduplex circles complementary to those illustrated in the figure. In scheme *a*, for example, the plus-strand of p255/2 and the minus-strand of p255/11 would form a heteroduplex, similar in structure to the one illustrated for the minus-strand of p255/2 and the plus-strand of p255/11. For reasons of fiscal economy, I did not include in the reaction the oligonucleotide required to complement and mutagenize the gapped sequence in the 'complementary' circle, which therefore regenerated parental-type DNA. Despite the high background of parental-type plasmids, the desired mutants were readily identified by the loss of the *Hin*dIII restriction site preceding the ATG codon, or by the gain of an *Ava*II cleavage site following the ATG triplet. The protocol outlined in scheme *a* gave rise directly to mutants designated p255/15 and p255/16, which are identical to p255/2 (described in Fig. 1) except for the nucleotide in the −3 position. From p255/15 and p255/16, derivatives p255A/3 and p255A/4 were subsequently obtained by fusing the *Bcl*I site to the *Bam*HI site, as illustrated for p255A/1 in Fig. 1*c*.
Methods: Linear DNA from two plasmids (identified in the diagrams), each cleaved with a different single-cutting restriction enzyme, was denatured by boiling for 1 min and then annealed (at 50 μg ml^{-1}) overnight in 20 mM Tris-HCl (*p*H 7.7), 150 mM NaCl and 0.2 mM EDTA, in a water bath that was gradually cooled from 85 °C to room temperature. The reannealed DNA was concentrated by ethanol precipitation and redissolved in a small volume of buffer. 1 μg of the DNA was mixed with 0.02 A_{260} units of the 5′-phosphorylated oligonucleotide that is shown, in the diagram, slightly to the right of the heteroduplex. The oligonucleotides used in schemes *a* and *b* were custom-synthesized by PL Biochemicals, Inc. The DNA/oligonucleotide mixture in 100 μl of 50 mM Tris-HCl (*p*H 7.6), 10 mM $MgCl_2$ and 150 mM NaCl was incubated at 65 °C for 5 min and then held at 15 °C for 16 h. T_4 DNA ligase (4×10^3 units ml^{-1}, New England BioLabs) was present during the 15 °C incubation, together with 1 mM ATP, 20 mM dithiothreitol and bovine serum albumin at 50 μg ml^{-1}. The ligase reaction was stopped by heating at 65 °C for 10 min. An aliquot of the reaction mixture was used directly to transform *E. coli* MM294, according to standard procedures[23]. Ampicillin-resistant colonies were picked and the DNA from 10-ml minicultures[24] was screened with *Hin*dIII or *Ava*II to identify the desired mutants.

The yield of proinsulin varied with each of the mutants tested in Fig. 4. Comparison of lanes 1–3 shows that p255A/3 synthesized ~15-fold more proinsulin than did p255A/1; the translational efficiency of p255A/4 fell between those two extremes. To control for possible variations in transfection efficiency from plate to plate, each cell monolayer was transfected in this experiment with a mixture of two plasmids: the mutant plasmid under investigation (for example, p255A/1) and a previously characterized derivative called p255/10. The salient feature of p255/10 is that it directs synthesis of an elongated form of preproinsulin, as shown in lane 6 of Fig. 4. Using the high molecular weight insulin-related polypeptide encoded by p255/10 as an internal control, the amount of proinsulin produced by p255A/1 and each of its derivatives can be reliably compared. The data in Fig. 4 indicate that the mutant plasmids direct synthesis of proinsulin with vastly different efficiencies. Table 1 summarizes the sequences and the relative translational efficiencies of the various plasmids in the A-series.

The targeted mutagenesis technique used in this study is unlikely to produce secondary mutations outside the region of interest. Although the oligonucleotides used for mutagenesis are rather short, the structure of the gapped heteroduplex circle precludes binding of the oligonucleotides to extraneous sites. Nevertheless, it seemed desirable to prove that the observed variations in translation were due solely to sequence differences around the ATG initiator codon. Towards that end, I subjected the low-yielding plasmid p255A/2 to a further mutagenesis step, cleaving the DNA first with *Hin*dIII and then using S_1

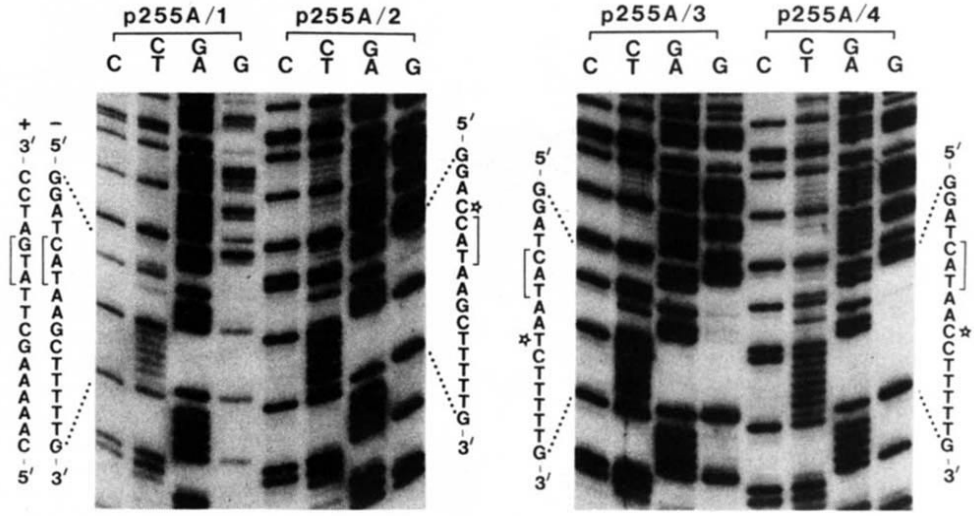

Fig. 3 Sequence analysis of mutant plasmids. The gels show the sequence of the minus-strand of a portion of the DNA near the start site for translation. The first mutant in this series, p255A/1, is shown at the far left; for convenience, a portion of the plus-strand sequence of p255A/1 is written beside the minus-strand. In the sequences of the other mutants, the nucleotides that differ from p255A/1 are starred. The ATG triplet or its complement is bracketed. **Methods:** The 250-bp fragment used for sequencing extends from the *Dde*I site in SV40 DNA (56 bp upstream from the unique *Hin*dIII site shown in Fig. 1*b*) to an *Eco*RI site in the large intron of the rat insulin gene. After cutting the plasmid DNA with *Dde*I, the indented 3′ termini were labelled by incubating with TTP, [α-^{32}P]dCTP (800 Ci mmol^{-1}; NEN) and *E. coli* DNA polymerase I (Klenow fragment, New England BioLabs). The fragment of interest was recut with *Eco*RI and purified by electrophoresis through a 5% polyacrylamide gel. Sequencing was carried out by the chemical cleavage method of Maxam and Gilbert[25] using 8% polyacrylamide gels. The gels were dried before autoradiography.

nuclease to delete five nucleotides directly in front of the ATG codon; the new derivative is called p255A/6. The yield of proinsulin from p255A/6 was at least 15-fold greater than from p255A/2. Thus, inefficient expression of proinsulin by p255A/2 can be attributed to the sequence context around the initiator codon rather than to a spurious second mutation elsewhere in the cloned gene.

Discussion

The experiments presented above provide the first evidence that sequence context influences the ability of an AUG triplet to be recognized as an initiator codon by eukaryotic ribosomes. p255A/1 and p255A/3 provide the most dramatic contrast: changing the nucleotide in the −3 position from C to A enhanced translation of rat proinsulin at least 15-fold. The approximately three-fold difference in translation observed between p255A/3 and p255A/4 indicates that A functions better than G in the −3 position, although either purine in that position is far superior to cytidine. The contribution of the nucleotide in the +4 position is less clear, since p255A/4 and p255A/7 differed by at most 1.5-fold in their production of proinsulin. Attempts are underway to replace the purine in the +4 position with a pyrimidine, which might have more dramatic consequences.

Although I have tested so far only the −3 and +4 positions, which are the most highly conserved positions in natural mRNAs[10], it seems likely that other nearby nucleotides also contribute. I have detected measurable differences in translational efficiency among a number of synthetic initiation sites, all of which have A in the −3 position but differ in the adjacent nucleotides (M. K., unpublished data). Obviously a more extensive set of systematically derived point mutants is needed to determine the full extent of the flanking region that contributes to a 'favourable' initiation site. It also seems likely that structural features in mRNA apart from the environs of the AUG triplet can influence translational efficiency. Thus, the prediction that emerges from the present study must be carefully circumscribed: all other things being equal, translational efficiency should increase dramatically when a purine replaces a pyrimidine in the −3 position; and A in that position works considerably better than G. However, the experiments described here do not predict that any two messages with identical sequences flanking the initiator codon will necessarily be translated with identical efficiencies, since some other feature (such as the sequence adjacent to the cap, or the secondary structure of the message, or the pattern of codon usage) might differ in a way that affects messenger function. Defining those ancillary features is a task for the future.

Additional experiments are also needed to determine precisely what happens when the migrating 40S ribosomal subunit encounters an AUG triplet in an unfavourable sequence context—which, for the time being, can be taken to mean a C in the −3 position. One possibility is that most of the 40S subunits detach from the message. Abortive initiation with release of 40S subunits would account for the observed low yield of proinsulin from p255A/1 and p255A/2. An alternative mechanism is that 40S subunits remain bound to the message and simply bypass an AUG triplet that occurs in an unfavourable context. To evaluate these two possibilities (which are not mutually exclusive), I have constructed plasmids with an ATG triplet upstream and in a different reading frame from the preproinsulin start site, and have monitored the yield of proinsulin as a function of varying the sequence around the upstream ATG triplet. Those experiments are described elsewhere[17].

Inspection of the data in Fig. 4 reveals that p255A/3, which is the best of the mutants analysed in this series, directs synthesis

Table 1 Relative translational efficiency of A-series mutants

Plasmid	Sequence at start of preproinsulin coding region (−3 +4)	Relative yield of proinsulin
p255A/1	CAAAAAGCTTATGATCCG	1
p255A/4	CAAAAAGG*TTATGATCCG	5
p255A/3	CAAAAAGA*TTATGATCCG	15
p255A/2	CAAAAAGCTTATGG*TCCG	1±1.5
p255A/7	CAAAAAGG*TTATGG*TCCG	5–7.5

The underlined sequence TGATCC in p255A/1 is the site of the *Bcl*I–*Bam*HI fusion described in Fig. 1 legend. The *Hin*dIII recognition site (AAGCTT) occurs just upstream from the ATG triplet in p255A/1. In the sequences shown for the other mutants, the nucleotides that differ from p255A/1 are starred. The yield of proinsulin from p255A/1 has been set arbitrarily at 1. The relative yield of proinsulin was determined by running, on a single gel, serial dilutions of each sample. All samples could thus be quantitated from a single autoradiogram. This gave more reproducible results than I was able to obtain by applying the samples (undiluted) to a gel and then comparing a series of autoradiograms exposed for different lengths of time.

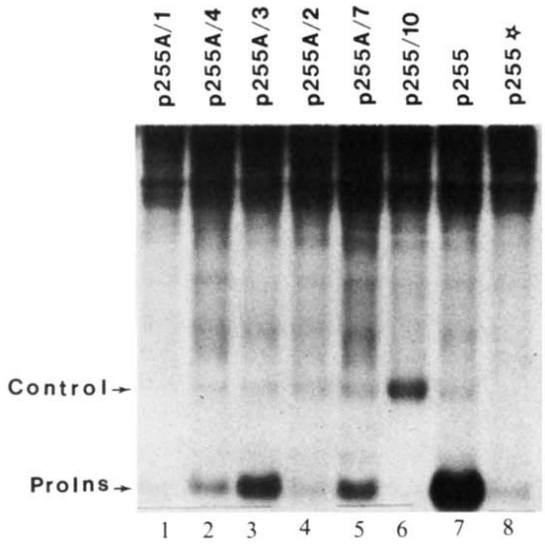

Fig. 4 Variations in proinsulin synthesis as a function of sequence context around the AUG initiator codon. Lanes 1–5 show the yield of proinsulin from cells transfected by mutants in the A-series. For comparison, proinsulin synthesis in cells transfected by wild-type p255 is also shown (lane 7). The proinsulin band is marked 'ProIns', and the high molecular weight insulin-related polypeptide encoded by p255/10 (see text) is marked 'Control'. The intensity of the control band is lower in other lanes, relative to lane 6, due to competition by the test plasmids. The important point, however, is that the intensity of the control band is constant in lanes 1–5 (this was confirmed after prolonged exposure of the autoradiogram), whereas the yield of normal-sized proinsulin varies. For the control shown in lane 8, excess unlabelled insulin was included during the incubation with antiserum, preventing precipitation of both proinsulin and the high molecular weight band encoded by p255/10.

Methods: COS-1 cells were transfected[8,13] with a mixture of the control plasmid p255/10 (5 μg per plate) and 20 μg of one of the test plasmids indicated across the top of the figure. The sample in lane 6 is from cells transfected only with p255/10. Forty-eight h after DNA uptake, the cells were labelled for 3.5 h with ^{35}S-cysteine (NEN, 1,000 Ci mmol^{-1}). Subsequent extraction and immunoprecipitation of the labelled proteins were carried out as described previously[8]. The polypeptides were fractionated on a 15% polyacrylamide-SDS gel[26] containing 6 M urea.

of significantly less proinsulin than does wild-type p255. Several factors might account for this: (1) The polypeptide encoded by p255A/3 and the other plasmids in the A-series lacks three amino acids that are present at the amino-terminus of the natural form of rat preproinsulin. This difference in structure might affect the intracellular processing or stability of the polypeptide. The difference in proinsulin yield between wild type and p255A/3 persists, however, even when the proteins are analysed after a very short pulse with ^{35}S-cysteine. (2) The production and/or stability of mature mRNA might be lower in the case of p255A/3 than in the case of wild-type p255. This explanation seems unlikely, since Lomedico and McAndrew[11] detected no transcriptional anomalies when they perturbed the 5′ structure of the rat preproinsulin gene. (3) Although both the wild-type preproinsulin gene carried by p255 and the mutant form carried by p255A/3 have an A residue three nucleotides upstream from the ATG initiator codon, the sequences around the initiator codons in the two plasmids are otherwise quite dissimilar: the wild-type sequence is CCAACATGG, versus AGATTATGA for p255A/3. I suspect that this difference underlies the more efficient expression of the wild-type preproinsulin gene; but that suspicion awaits verification. It is important to recognize that the conclusions drawn from this study derive not from comparing the wild-type gene with any of the mutants, but from comparing one mutant with another within the A series.

The finding that nucleotides flanking the AUG initiator codon modulate translational efficiency is not unexpected. Although the scanning model has always emphasized the importance of position (that is, proximity to the 5′ end of the message) in defining the functional initiation site, the model was recently revised in the light of two observations. (1) Although translation usually begins at the 5′-proximal AUG triplet in eukaryotic mRNAs, there is a growing list of exceptional cases (~5% of cellular mRNAs and ~10% of viral mRNAs) in which one or more AUG triplets occur some distance upstream from the start of the protein coding sequence[9,10,18]. Thus, eukaryotic ribosomes do not always initiate at the first AUG triplet in the message. (2) The distribution of nucleotides flanking functional initiator codons in eukaryotic mRNAs shows a striking bias in position −3 (where A occurs in ~80% of the messages examined) and in position +4 (where G occurs in 40–60%). The nucleotide preference in positions −1, −2, −4 and −5 is less striking, but a predominance of C has been noted in those positions[9,10]. The nonrandom distribution of nucleotides suggests that the initiation machinery is not indifferent to the sequence context around the AUG triplet. Moreover, the nucleotides—most notably, a purine in the −3 position—that distinguish functional initiator codons rarely occur around the 'nonfunctional' upstream AUG triplets in the exceptional mRNAs cited above. These observations were incorporated into a modified scanning model that emphasizes the importance of both position and sequence context in identifying the initiation site. The modified scanning model[10,18] states that a 40S ribosomal subunit binds initially at the 5′ end of mRNA and subsequently migrates until it reaches the first AUG triplet: if the first AUG triplet occurs in an optimal sequence context, all 40S subunits stop there and that AUG serves as the unique initiator codon, but if the first AUG triplet occurs in a suboptimal sequence context, only some 40S subunits stop and initiate there; some bypass that site and initiate at another AUG that lies farther downstream. The experiments described here represent the first step towards defining the 'optimal sequence context' for initiation by eukaryotic ribosomes.

I thank Peter Lomedico for providing the parental plasmid p255. This research was supported by a grant from the NIH. M.K. is the recipient of Research Career Development Award AI00380.

Received 22 September; accepted 1 December 1983.

1. Kozak, M. *Cell* **15,** 1109–1123 (1978); **22,** 7–8 (1980); *Curr. Topics Microbiol. Immun.* **93,** 81–123 (1981).
2. Kozak, M. *Nature* **280,** 82–85 (1979).
3. Kozak, M. *Cell* **22,** 459–467 (1980).
4. Kozak, M. & Shatkin, A. J. *J. biol. Chem.* **253,** 6568–6577 (1978).
5. Privalsky, M. L. & Bishop, J. M. *Proc. natn. Acad. Sci. U.S.A.* **79,** 3958–3962 (1982).
6. Perdue, M. L., Borchelt, D. & Resnick, R. J. *J. biol. Chem.* **257,** 6551–6555 (1982).
7. Preston, C. M. & McGeoch, D. J. *J. Virol.* **38,** 593–605 (1981).
8. Kozak, M. *Cell* **34,** 971–978 (1983).
9. Kozak, M. *Nucleic Acids Res.* (in the press).
10. Kozak, M. *Nucleic Acids Res.* **9,** 5233–5252 (1981).
11. Lomedico, P. T. & McAndrew, S. J. *Nature* **299,** 221–226 (1982).
12. Gluzman, Y. *Cell* **23,** 175–182 (1981).
13. Wigler, M., Pellicer, A., Silverstein, S. & Axel, R. *Cell* **14,** 725–731 (1978).
14. Lomedico, P. T. *Proc. natn. Acad. Sci. U.S.A.* **79,** 5798–5802 (1982).
15. Gruss, P. & Khoury, G. *Proc. natn. Acad. Sci. U.S.A.* **78,** 133–137 (1981).
16. Laub, O. & Rutter, W. J. *J. biol. Chem.* **258,** 6043–6050 (1983).
17. Kozak, M. *Nucleic Acids Res.* (submitted).
18. Kozak, M. *Microbiol. Rev.* **47,** 1–45 (1983).
19. Kalderon, D., Oostra, B. A., Ely, B. K. & Smith, A. E. *Nucleic Acids Res.* **10,** 5161–5171 (1982).
20. Oka, A., Sugimoto, K., Sasaki, H. & Takanami, M. *Gene* **19,** 59–69 (1982).
21. Peden, K. W. C. & Nathans, D. *Proc. natn. Acad. Sci. U.S.A.* **79,** 7214–7217 (1982).
22. Kramer, W., Schughart, K. & Fritz, H-J. *Nucleic Acids Res.* **10,** 6475–6485 (1982).
23. Morrison, D. A. *Meth. Enzym.* **68,** 326–331 (1979).
24. Klein, R. D., Selsing, E. & Wells, R. D. *Plasmid* **3,** 88–91 (1980).
25. Maxam, A. M. & Gilbert, W. *Meth. Enzym.* **65,** 499–560 (1980).
26. Laemmli, U. K. *Nature* **227,** 680–685 (1970).