Cell, Vol. 44, 283–292, January 31, 1986, Copyright © 1986 by Cell Press

Point Mutations Define a Sequence Flanking the AUG Initiator Codon That Modulates Translation by Eukaryotic Ribosomes

Marilyn Kozak
Department of Biological Sciences
University of Pittsburgh
Pittsburgh, Pennsylvania 15260

Summary

By analyzing the effects of single base substitutions around the ATG initiator codon in a cloned preproinsulin gene, I have identified ACCATGG as the optimal sequence for initiation by eukaryotic ribosomes. Mutations within that sequence modulate the yield of proinsulin over a 20-fold range. A purine in position −3 (i.e., 3 nucleotides upstream from the ATG codon) has a dominant effect; when a pyrimidine replaces the purine in position −3, translation becomes more sensitive to changes in positions −1, −2, and +4. Single base substitutions around an upstream, out-of-frame ATG codon affect the efficiency with which it acts as a barrier to initiating at the downstream start site for preproinsulin. The optimal sequence for initiation defined by mutagenesis is identical to the consensus sequence that emerged previously from surveys of translational start sites in eukaryotic mRNAs. The mechanism by which nucleotides flanking the ATG codon might exert their effect is discussed.

Introduction

A considerable body of evidence supports the idea that 40S ribosomal subunits bind at the capped 5′ end and scan the mRNA sequence until an AUG codon is reached (Kozak, 1980a, 1983a). The fact that 40S subunits can migrate on mRNA, prior to the assembly of a complete 80S ribosome, has been demonstrated in vitro (Kozak and Shatkin, 1978; Kozak, 1980b). The stimulatory effect of the m^7G cap (Shatkin, 1976) and the inability of ribosomes to bind to circular mRNAs (Kozak, 1979; Konarska et al., 1981) point to an end-dependent mechanism. The fact that eukaryotic ribosomes usually translate only the 5′-proximal cistron in polycistronic viral mRNAs (Smith, 1977) is also rationalized by the scanning model. The importance of position in defining the functional initiation site was shown by manipulating a cloned preproinsulin gene to produce an mRNA in which the "ribosome binding site" (i.e. the ATG initiator codon and flanking sequence) was tandemly reiterated: ribosomes initiated exclusively at the 5′-proximal copy in the tandem array (Kozak, 1983b).

Inspection of sequences near the 5′ ends of eukaryotic mRNAs provides evidence that might be interpreted for or against the scanning hypothesis: in ~90% of the mRNAs examined, there are no extraneous AUG triplets upstream of the functional initiator codon—a provocative finding that is rationalized by the scanning model. However, 5% to 10% of eukaryotic mRNAs have AUG triplet(s) upstream of the known start site for protein synthesis (Kozak, 1983a). In such mRNAs, the upstream AUG triplets occur in a context different from the conserved pattern of nucleotides around functional initiator codons (Kozak, 1981, 1983a, 1984a). This difference inspired a modified version of the scanning model in which both the position of an AUG codon and its context play a role (Kozak, 1981). Our current working hypothesis is that a 40S ribosomal subunit (with associated factors, of course) binds at the 5′ end of mRNA and advances linearly until it reaches the first AUG triplet: if the first AUG codon occurs in an optimal context, all 40S subunits stop and that AUG serves as the unique site of initiation. But if the sequence around the first AUG triplet is suboptimal, some 40S subunits bypass that site and initiate farther downstream. The optimal context for initiation, derived from the aforementioned survey, was $CC^A_GCCAUGG$ (Kozak, 1981, 1984a). Within that sequence, the purine in position −3 (3 nucleotides upstream of the AUG codon) is most highly conserved: ~75% of the mRNAs examined had A in that position, and another 20% had G. Some experimental evidence for the importance of A or G in position −3, and G in position +4 (immediately following the AUG codon), has been obtained by measuring the binding of synthetic oligonucleotides to wheat germ ribosomes in vitro (Kozak, 1981). By applying site-directed mutagenesis to a cloned preproinsulin gene, point mutations were created near the AUG initiator codon; translation of those mutants in vivo confirmed the requirement for a purine in position −3 (Kozak, 1984b). Using a more efficient scheme for mutagenesis, I have now obtained a larger set of mutants. Base substitutions in at least four positions near the AUG codon modulate its function, as described below. The sequence ACCAUGG emerges as the best context for initiation in this system. Morlé et al. (1985) recently described an α-thalassemia in which the sequence at the initiation site for α-globin was changed from CACCAUG to CCCCAUG. The resulting deficiency in globin synthesis confirms, in a natural setting, the importance of A in position −3.

There are a few eukaryotic mRNAs in which an AUG codon in an excellent context (such that all 40S subunits should initiate exclusively there) occurs upstream of a second AUG codon which is known to function. The scanning mechanism as outlined above cannot explain that anomaly. Initiation at the internal AUG triplet in such messages has been attributed to reinitiation, which seems reasonable because access to the downstream cistron is critically dependent on having a terminator codon in frame with the first AUG codon and upstream from the second (Dixon and Hohn, 1984; Hughes et al., 1984; Kozak, 1984c; Liu et al., 1984). Our interpretation is that the 40S ribosomal subunit remains on the mRNA at the terminator codon and resumes scanning; it stops and reinitiates when it reaches the next (internal) AUG codon. The similarity between primary initiation (i.e. selection of the 5′-proximal AUG codon) and reinitiation is emphasized by the finding, described herein, that the optimal context for the AUG codon is the same for both processes.

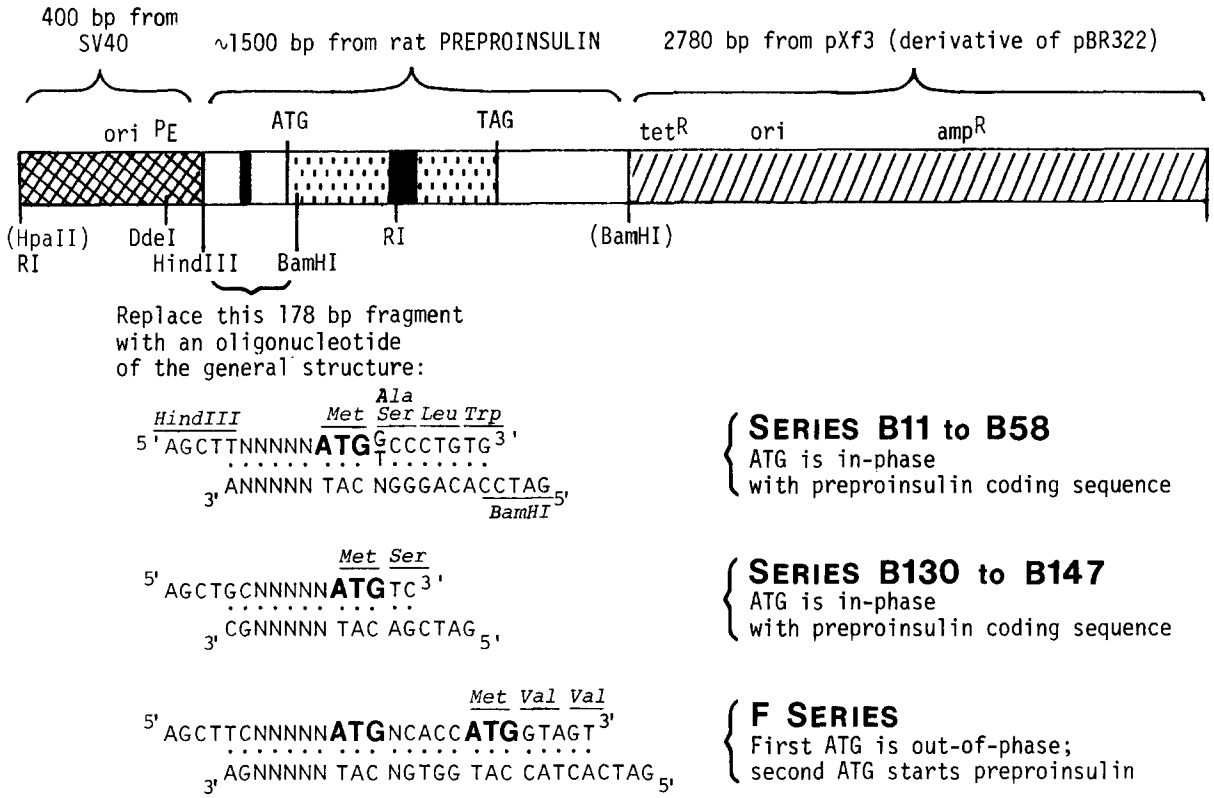

Figure 1. Scheme for Generating Point Mutations around the ATG Codon in Plasmids That Express Preproinsulin

The parental plasmid p255/11 is shown in linear form at the top. Transcription of the rat preproinsulin gene is mediated by the SV40 early promoter, designated P_E. Shaded areas represent introns that interrupt the 5'-noncoding sequence (open areas) and the preproinsulin coding sequence (stippled). The Bam HI site shown in parentheses is present in the original p255 (Lomedico and McAndrew, 1982) but was eliminated from p255/11 (see Experimental Procedures). "N" indicates positions where the sequence of the oligonucleotides was ambiguous, generating mutations around the ATG codon. The general form of the mutagenic insertion sequences is shown here; the specific oligonucleotides used in the experiments are listed in Table 1. In the B series, mutations were introduced around an ATG codon that starts the preproinsulin coding sequence. In the F series, the focus of mutagenesis was an upstream, out-of-frame ATG codon that serves as a barrier to initiating preproinsulin. The N-terminal amino acid sequence is shown for the variant form of preproinsulin encoded in each series. In the case of B11 through B58, the encoded protein is identical with wild-type rat preproinsulin (MetAlaLeuTrp).

Results

Mutagenesis around the Initiator Codon in Preproinsulin Expression Vectors

Figure 1 shows the scheme used to produce mutants in the B series, in which the sequence flanking the initiator codon was systematically varied. The technique involves deleting the DNA fragment that carries the normal translational start site for preproinsulin, and replacing it with a synthetic ATG-containing oligonucleotide: the ATG codon carried on the insert functions as the new initiation site for preproinsulin. (Since all manipulations in this study were at the DNA level, I shall refer to the initiator codon as ATG irrespective of whether the reference is to DNA or mRNA.) The presence of sticky ends on the oligonucleotide that are complementary to those on the acceptor DNA ensures efficient insertion, and the presence of sequence ambiguities within the oligonucleotide generates a large number of mutants. Matteucci and Heynecker (1983) were the first to use a technique similar to this for analyzing translation in E. coli.

Multiple insertions were precluded by carrying out the ligase reaction without prior 5' phosphorylation of the oligonucleotide; thus, each recovered mutant carried a single copy of the oligonucleotide. Only when the mismatch occurred very close to the end of the oligonucleotide, in the penultimate or antepenultimate position, did I encounter difficulty in obtaining the nearly complete set of mutants expected. Occasionally, however, one member of a series was not found, despite the repeated isolation of other members of the set; in such cases, a new oligonucleotide was prepared that encoded unambiguously the desired sequence. Mutants in which the sequence of the insert did not correspond precisely to the starting oligonucleotide were recovered infrequently; when found, they usually deviated in one position from the input oligonucleotide. No mutations were found outside of the region encompassed by the oligonucleotide insert.

The screening of mutants was facilitated by varying only two or three positions at a time. Thus, three oligonucleotides, each of which conforms to the general structure shown in Figure 1, were actually used to obtain mutants

Table 1. Oligonucleotides Used for Insertion Mutagenesis

Mutants	Oligonucleotides
B11 - B21	AGCTTCCANNATGGCCCTGTG
B31 - B39	AGCTTGGNTTATG$^{G}_{T}$CCCTGTG
B41 - B58	AGCTTGG$^{A}_{G}$NNATGTCCCTGTG
B130 - B135	AGCTGCC$^{AAA}_{CCC}$CATGTC
B137 - B138	AGCTGCTT$^{A}_{C}$TTATGTC
B140 - B143	AGCTGCT$^{A}_{T}$T$^{A}_{T}$TATGTC
B145 - B147	AGCTGC$^{TT}_{CC}$T$^{TT}_{CC}$ATGTC
F1 - F8	AGCTTCTG$^{A}_{G}$TTATGNCACCATGGTAGT
F9 - F10	AGCTTCTG$^{A}_{T}$TT$^{A}_{C}$TGTCACCATGGTAGT
F11 - F15	AGCTTCC$^{C}_{G}$A$^{TT}_{CC}$ATGTCACCATGGTAGT

Sequences are listed for the top (plus) strand of each duplex oligonucleotide that was inserted between the Hind III and Bam HI sites of p255/11. "N" stands for a mixture of all four nucleotides. The mutants that were obtained are listed in the left column; their general form is depicted in Figure 1, and the precise sequence around the ATG codon in each mutant is given in Figures 2 through 6.

B11 through B58. The sequences of the specific oligonucleotides used for mutagenesis are given in Table 1. In the case of B11 through B58, the protein initiated at the inserted ATG codon has the same N-terminal amino acid sequence as wild-type preproinsulin. For reasons of economy, shorter oligonucleotides were used to obtain mutants B130 through B147, and the first few amino acids of the wild-type protein were not retained. This does not compromise the results that follow because the yield of proinsulin from each mutant was always compared, not to the wild-type plasmid, but to a control from the same series as the mutant.

To determine how single base changes around the initiator codon affect translational efficiency, the mutant plasmids were transfected into monkey (COS) cells as described in Experimental Procedures. Two days after transfection, the cells were incubated with ^{35}S-cysteine, and labeled proteins were extracted, immunoprecipitated, and analyzed by polyacrylamide gel electrophoresis. Because the primary translation product undergoes cleavage in these cells, the product that accumulates and was measured is proinsulin. Measurement of cytoplasmic RNA levels, as described in Experimental Procedures, revealed no significant differences among the mutants in a given series. Thus, the observed variation in proinsulin synthesis reflects the efficiency with which the mRNA produced by each plasmid is translated.

Effects of Mutating Positions −3 and +4

The sequences of mutants B31 through B39 are identical except for positions −3 and +4. Single nucleotide changes in those positions modulate the yield of proinsulin over a 20-fold range (Figure 2). Comparison of B35, B38, and B39 shows that A functions better than G, and G better than T, in position −3. Comparison of B38 with B34, or B39 with B33, reveals that G works better than T in position +4. The contributions of positions −3 and +4 are not simply additive. For example, G in position +4 enhanced translation about 5-fold with T in position −3 (B34 versus B38), 4-fold with G in position −3 (B33 versus B39), and only 2-fold with A in position −3 (B31 versus B35). Similarly, with mutants B31 through B34, where the favored nucleotide G occurs in position +4, the effects of varying position −3 were less dramatic than in mutants that had T in position +4. The hierarchy in position −3 (A > G > T) does not change, but the magnitude of the effect obtained upon mutating position −3 depends on how favorable the rest of the sequence is.

The data in Figure 2 confirm that proinsulin is quantitatively recovered in the first round of immunoprecipitation. The second cycle is not shown in the figures that follow.

Effects of Mutating Positions −1, −2, −4, and −5

Mutagenesis was limited to positions −1 and −2 in the first experiment. As shown in Figure 3, C in both of those positions enhanced translation marginally, at best. The bracketed lanes in that figure represent duplicate plates that were transfected with the same plasmid, establishing that the variability of the assay is ≤20%. With that in mind, it seems safe to draw conclusions about plasmids that differ 2-fold or more in their production of proinsulin; but when the increment is less than 2-fold, as with B41, the result cannot be considered more than suggestive. The conclusion from Figure 3 is that, if the nucleotides in positions −1 and −2 influence translation at all, their contributions are small.

To pursue the issue, I constructed the mutants shown in Figure 4. The 3-fold decrease in the yield of proinsulin between B137 and B138 confirms that A works better than C in position −3, as we already knew. The 3-fold increase in proinsulin between B138 and B130 provided the first evidence that C in (some or all of) positions −1, −2, −4, and −5 enhances translation. The effect can also be seen with mutants that have A in position −3 (B137 versus B133), although the stimulation was only 2-fold. (As noted above, nonadditivity can minimize the effects of some sequence changes. Indeed, introducing C into positions −1, −2, −4, and −5 was without measurable effect in mutants that had the optimal A in position −3 plus G in position +4 [B11 through B21, data not shown].) The mutants studied in Figure 4 provide two peripheral insights: C in position −3 apparently functions better than T (B138 versus B140); and, in mutants that lack A in position −3, the presence of A in position −2 or −4 does not compensate. Whereas A in position −3 stimulates translation about 10-fold (B137 versus B140), mutants B141 and B143 translate only marginally better than B140.

Since the stimulatory effect of Cs in positions −1, −2, −4, and −5 was more apparent in plasmids that had C rather than A in position −3 (Figure 4), and since T_{-3} seemed to be even weaker than C_{-3}, it seemed that plasmids with T in position −3 might provide the most sensitive background for further studies. Accordingly, the mutants shown in Figure 5 were constructed and analyzed. Introducing Cs into positions −1 and −2 indeed stimulated translation at least 4-fold (B145 versus B140). Surprisingly, however, Cs in positions −4 and −5 did not

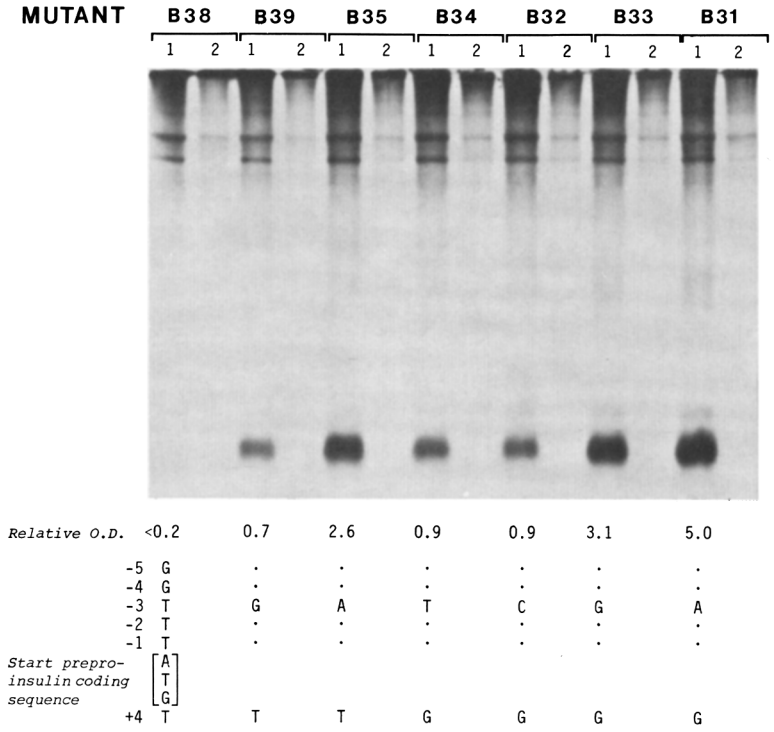

Figure 2. Variation in Proinsulin Synthesis Caused by Single Base Changes in Positions −3 and +4

^{35}S-labeled proteins from transfected COS cells were subjected to two rounds of immunoprecipitation which were analyzed in contiguous lanes of the gel, marked 1 and 2. An arrow indicates the position of proinsulin. To make the figure easily readable, only the nucleotides that differ among members of the set are indicated. The sequence shown for B38 in positions −1, −2, −4, and −5 is common to the whole set.

enhance translation (B146 versus B140). The yield of proinsulin from B147, where Cs occur in all four positions, was no greater than from B145.

The 20-fold variation in proinsulin synthesis among mutants in the B series confirms that sequences flanking the ATG codon modulate translational efficiency, but the B mutants do not reveal what happens at a weak ATG codon. The scanning model predicts that some 40S subunits bypass an ATG codon that occurs in an unfavorable context, and translation begins farther downstream. To test that prediction, mutants were needed in which initiation at a downstream ATG codon could be monitored. The F series described below meets that requirement.

Effects of Single Base Substitutions around an Upstream ATG Codon

The general form of the oligonucleotides that were inserted to generate mutants in the F series is shown in Figure 1. The position of the second ATG triplet carried on the insert enables it to serve as the initiator codon for preproinsulin, and the sequence around that ATG codon was not mutated. Rather, point mutations were introduced around an upstream, out-of-frame ATG triplet that was expected to function as a "barrier," reducing the number of 40S ribosomes that reach the preproinsulin start site. For this design to work, it was important that no terminator codons occur between the upstream ATG triplet and the preproinsulin start site: eukaryotic ribosomes can reinitiate at the second ATG codon when an upstream "minicistron" terminates prior to the second ATG codon (see Introduction), and the inhibitory effect of an upstream ATG

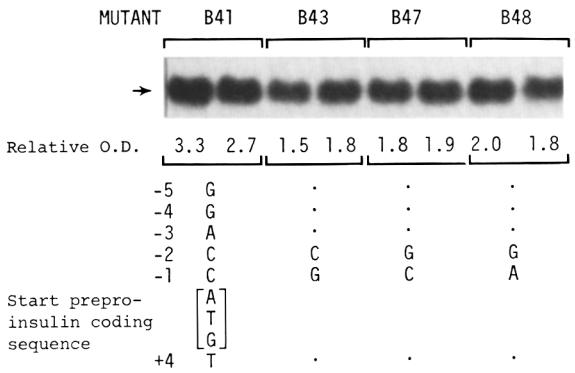

Figure 3. Proinsulin Synthesis by Plasmids That Vary in Positions −1 and −2

Adjacent lanes enclosed by brackets in this and subsequent figures represent ^{35}S-labeled proteins from duplicate plates that were transfected with the same plasmid. The fluorograms have been cropped to show only the proinsulin region of the gel, which is marked by an arrow. The sequence that is shown for B41 in positions −3, −4, −5, and +4 is common to all four plasmids used for this experiment.

barrier is thereby reversed. To preclude reinitiation in the F series, the reading frame established by the upstream ATG codon had to extend beyond the preproinsulin start site; but the ideal site for termination was hard to determine. When the upstream cistron is long, i.e. when it overlaps the preproinsulin coding sequence over a considerable distance, there is interference (I think at the level of elongation) such that the yield of proinsulin is low even when the upstream ATG triplet occurs in a weak context

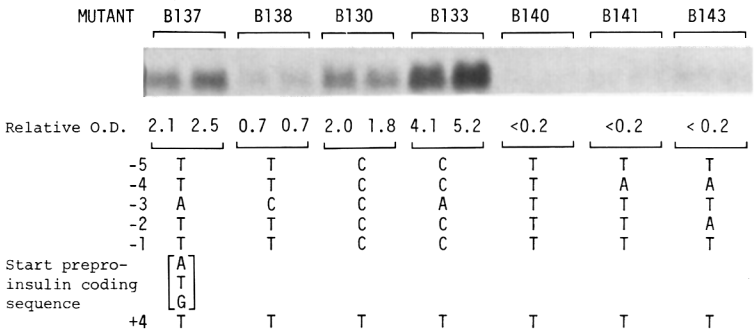

Figure 4. Expanded Mutagenesis of Sequences Preceding the ATG Codon

Details are given in the legend to Figure 3.

(Kozak, 1984c). Since that unexplained interference decreases as the extent of overlap between the two cistrons decreases, it might seem ideal for the codon that terminates the upstream cistron to overlap the ATG codon that initiates preproinsulin. Unfortunately, that overlapping configuration still allows reinitiation, albeit very inefficiently. The sequence ATGGTAG, which I ultimately chose for the F series, allows about 2% reinitiation (based on unpublished experiments with other mutants). It was a workable compromise: the potential sources of interference outlined above were sufficiently reduced that I could see systematic effects of context on the function of upstream ATG codons.

The mutants listed in Figure 6 were studied to determine how single base substitutions around the first ATG codon affect the ability of ribosomes to reach the second, where preproinsulin initiates. F10, which lacks an upstream ATG triplet, is the control to which all other F mutants are compared. In F9, the upstream ATG codon lies in a very weak context and, as expected, synthesis of proinsulin was only slightly reduced. As the context around the upstream ATG codon improves, it becomes a more effective barrier: synthesis of proinsulin dropped 5-fold with F6, F7, and F8, and 10-fold with F1 through F5. It is not surprising that F1—which has the optimal A in position −3 and G in position +4—still makes a trace of proinsulin. As shown in experiment 2 in Figure 6, the upstream barrier can be further strengthened by introducing Cs into the flanking positions. Moreover, because the structure of the F mutants allows a low level of reinitiation, synthesis of proinsulin cannot be shut off completely.

Whereas the results obtained with the B series allow one to conclude only that an ATG codon in one context works better than another, the results obtained with the F series justify the interpretation that 40S ribosomal subunits bypass an ATG codon that lies in a weak context.

Context Effects on Reinitiation

A 65 bp sequence that carries an ATG initiator codon, followed after 12 bp by a TAA terminator codon, was inserted at the Hind III site that lies just upstream of the preproinsulin start site in B34, B35, B39, and B38. Thus, ribosomes can make preproinsulin only by reinitiating at the second ATG codon in mutants B34R, B35R, B39R, and B38R. The yield of proinsulin from those plasmids varied about

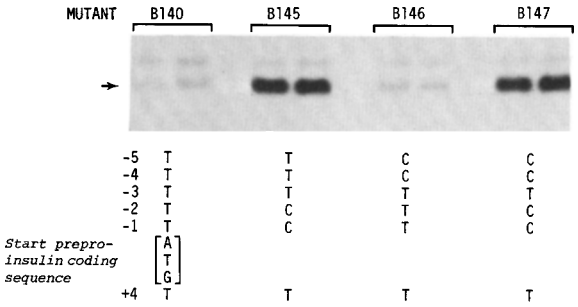

Figure 5. Translation Is Stimulated by C in Positions −1 and −2 When the Rest of the Sequence Is Suboptimal

Details are given in the legend to Figure 3.

8-fold, as shown in Figure 7. The "R" derivatives showed the same strong preference for a purine in position −3, and G in position +4, as was seen with the original B mutants. The control B37R lacks the second ATG codon and makes no proinsulin (Figure 7, bottom lane). This rules out the possibility that the proinsulin produced by the other "R" derivatives is actually initiated at the ATG codon carried on the insert (which would require suppression of the terminator codon and posttranslational cleavage to remove the N-terminal amino acid extension) rather than by reinitiating at the ATG codon that directly precedes the preproinsulin coding sequence.

Discussion

The Optimal Sequence for Initiation in Eukaryotes

From the foregoing mutational analysis, the sequence ACCATGG emerges as the most favorable context for initiation. Although I obtained no evidence that Cs in positions −4 and −5 are part of the ribosome recognition sequence, there is a lingering possibility that Cs in those positions contribute, but only in a small way, and perhaps only when a purine occurs in position −3. Inasmuch as Cs in those positions are highly conserved, I shall continue to show them, in parentheses, as part of the consensus sequence.

The validity of these experiments was ensured in several ways. In many experiments duplicate plates of COS cells were transfected with a given plasmid; the vari-

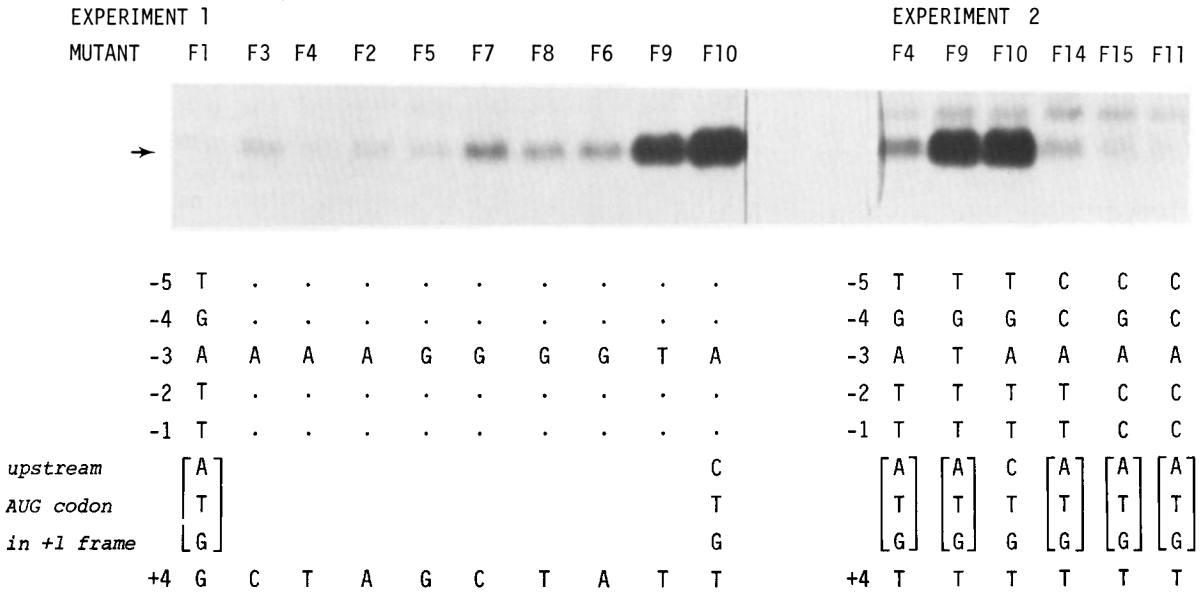

Figure 6. Effects of Context on the Efficiency of an Upstream ATG Barrier

The sequence shown for F1 in positions −1, −2, −4, and −5 is common to mutants F1 through F10. All have an upstream ATG codon, except for the control F10. The fluorograms have been cropped, and an arrow marks the position of proinsulin.

ation between duplicates never exceeded 20%. When too many plasmids had to be tested simultaneously, I did not assay in duplicate because it is hard to maintain precision when handling more than 12 or 14 plates; but subsequent repetitions of the experiment gave identical results. I considered one plasmid to be translated more efficiently than another only when the yield of proinsulin differed by at least 2-fold; often the difference was 5 to 20 fold. Every position in the consensus sequence has been tested in two or more independent constructs. For example, the combination of T in positions −3 and +4 was shown to be extremely weak in B38 and B140. Adenosine in position −3 worked best in B31, B35, B133, and B137. Comparison of B34 with B38, B33 with B39, and B31 with B35 confirms that G worked best in position +4. Comparison of B130 with B138, B133 with B137, and B147 with B140 confirms that C in some or all of positions −1, −2, −4, and −5 worked better than T. Finally, the results of the F series complement the B series: the sequence that gave the lowest yield of proinsulin when tested directly (T_{-3}/T_{+4} in B38) allowed the maximum production of proinsulin when that sequence was introduced as an upstream "barrier" in mutant F9.

Although the importance of A or G in position −3 was recognized easily and early (Kozak, 1984b), it was more difficult to show that other nearby nucleotides are recognized, because their contributions are not additive. The nonadditivity was an unwelcomed experimental complication, but it has a positive aspect in that most natural ribosome binding sites are buffered: given a purine in position −3, a mutation in position −1, −2, −4, or −5 should reduce translation only slightly. Thus, it is not surprising that mutations in those positions have not been described among the genetic diseases that have been studied at the molecular level. The dominant effect of position −3 in eukaryotic ribosome binding sites differs from the prokaryotic Shine-Dalgarno sequence, within which no single position is more important than any other. In striking contrast with the enhanced translation that occurred when A was introduced into position −3, there was little stimulation when A occurred in position −2 or −4 (compare B140 with B143 in Figure 4). Thus, if two separate components on the ribosome are responsible for recognizing the upstream A residue and the ATG codon, respectively, the two components must be rigidly oriented. This inflexibility with respect to the position of the upstream recognition sequence again differs from prokaryotes, where the distance between the initiator codon and the Shine-Dalgarno sequence is permitted to vary over a limited range (Kozak, 1983a).

The optimal context for recognition of the ATG codon appears to be the same for reinitiation at an internal site as for primary initiation at the 5′-proximal ATG codon. Our understanding of reinitiation is admittedly primitive, but a likely scenario is that, when an 80S ribosome reaches a terminator codon (having just translated a cistron or minicistron near the 5′ end of the mRNA), the 60S subunit detaches while the 40S subunit remains bound to the message and resumes scanning; when the 40S subunit reaches the next ATG codon, it reinitiates translation. An alternative mechanism postulates that ribosomes can bind directly to the internal site. Evidence against direct binding has been adduced previously (Kozak, 1984c). The notion of direct binding might have been revived had the present study revealed an optimal context for internal initiation different from the A_{-3}/G_{+4} motif that mediates recognition of the 5′-proximal ATG codon, Since the optimal context is the same for both processes, however, it seems

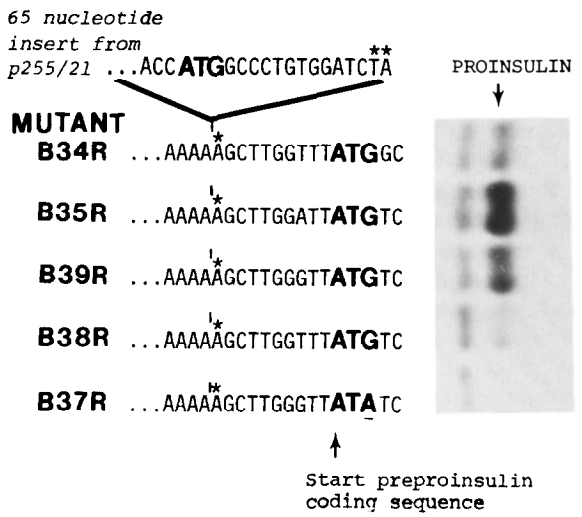

Figure 7. Single Nucleotide Changes in Positions −3 and +4 Affect the Efficiency of Reinitiation

In these "R" derivatives, ribosomes first initiate at the invariant ATG codon carried on the insert, terminate at the TAA codon (marked by asterisks), and then reinitiate 10 nucleotides downstream, at the start of the preproinsulin coding sequence, which differs (see positions −3 and +4) among the four mutants tested. B37R is a negative control described in the text. ^{35}S-labeled proteins from transfected COS cells were analyzed by polyacrylamide gel electrophoresis. The top of the gel (not shown) was at the left.

likely that reinitiation involves a replay of the scanning process.

Secondary Structure Does Not Contribute to the Observed Variation in Translation

Although secondary structure might differ slightly from one B mutant to another, that is unlikely to have influenced the results of this study. In an experiment described elsewhere (Kozak, 1986), I deliberately created a stable hairpin (ΔG −30kcal/mol) around the ATG codon in a mutant called B13hp; the hairpin did not reduce the yield of proinsulin. Thus 40S ribosomal subunits and/or the associated initiation factors have an impressive ability to melt duplex structures in mRNA. Since the point mutations described herein cannot create duplexes anywhere near as stable as the hairpin in B13hp, the 20-fold variation in proinsulin yield among members of the B series is best interpreted at the primary sequence level. The secondary structure hypothesis is also inconsistent with the dramatic inhibition that occurs upon changing a single nucleotide in position −3, compared with the small change in translation when several nearby nucleotides are mutated.

Incorporation of Context Rules into the Scanning Model

Our working hypothesis is that the migration of 40S subunits is halted more or less efficiently depending on the sequence around the ATG codon. This rationalizes the results obtained with the F series, where the strength of the upstream ATG "barrier" varies in the predicted way with changes in context. Constructs have been described previously in which the presence of an upstream ATG codon reduces or supplants initiation from the downstream site (Lomedico and McAndrew, 1982; Smith et al., 1983; Krieg et al., 1984) and, in some cases, the efficiency of the upstream ATG barrier was shown to be context dependent (Bandyopadhyay and Temin, 1984; Kozak, 1984c; Liu et al., 1984). The rules deduced by manipulating such cloned genes apparently hold for natural mRNAs as well—specifically for a number of viral mRNAs that have the unusual ability to produce two proteins. The first ATG codon in such mRNAs is usually in a weak context, thus rationalizing the ability of some ribosomes to reach the second ATG codon. There are at least ten examples from animal virus systems of bifunctional mRNAs that fit this pattern (Reddy et al., 1978; Bos et al., 1981; Clerx-van Haaster et al., 1982; Giorgi et al., 1983; Heermann et al., 1984; Laprevotte et al., 1984; Bellini et al., 1985; Castle et al., 1985; Clarke et al., 1985; Ernst and Shatkin, 1985; Persing et al., 1985; Sarkar et al., 1985). Such mRNAs in which there are two prominent initiation sites are rare. In most eukaryotic mRNAs the 5′-proximal ATG codon lies in a fairly strong context and ribosomes initiate predominantly at that site. To make more precise statements, we have to recognize that initiation sites are not simply strong or weak; the mutants described herein reveal a gradient of strength. The sequence (CC)ACC*ATG*G, ranks highest in efficiency: when that sequence occurs near the 5′ end of a message, no initiation can be detected downstream (see Kozak, 1983b; 1984c [mutants p255/20 and C2]; and mutant E13 in Kozak, 1986). A sequence such as AT*TATGT*, on the other hand, is also very efficient if one simply monitors the yield of protein initiated at that site (B35 in Figure 2). But in mutant F4, ~10% of the ribosomes bypassed the 5′-proximal AT*TATGT* and initiated at the preproinsulin start site downstream. If we now consider the sequences of natural eukaryotic mRNAs, although ~75% have A in position −3, only ~35% have A_{-3} plus G_{+4}; and less than 5% have the ideal (CC)ACC*ATG*G sequence. Thus, unless other features can compensate for a less than perfect context around the ATG codon, we should expect to find a second initiation site functioning (albeit very inefficiently) in most eukaryotic mRNAs. I suspect that compensatory mechanisms will be discovered. In some mRNAs where the coding sequence begins with a weak ATG codon, the objective might be not so much to allow ribosomes access to a second start site, but simply to limit the synthesis of a protein that would be harmful if overproduced. In the case of preproricin mRNA, the functional initiator codon is flanked by Ts in positions −3 and +4, and a strong out-of-frame ATG codon lies just upstream (Lamb et al., 1985). It would be hard to design a less favorable arrangement for translation—or a more toxic polypeptide.

By What Mechanism Do Nucleotides Flanking the ATG Codon Exert Their Effect?

Despite many differences between the prokaryotic and eukaryotic initiation mechanisms, the temptation to search 18S rRNA for a sequence that could do for eukaryotes what the Shine-Dalgarno sequence in 16S rRNA does for prokaryotes is irresistible. Figure 8 shows two sites in 18S rRNA that might pair with the (CC)ACC se-

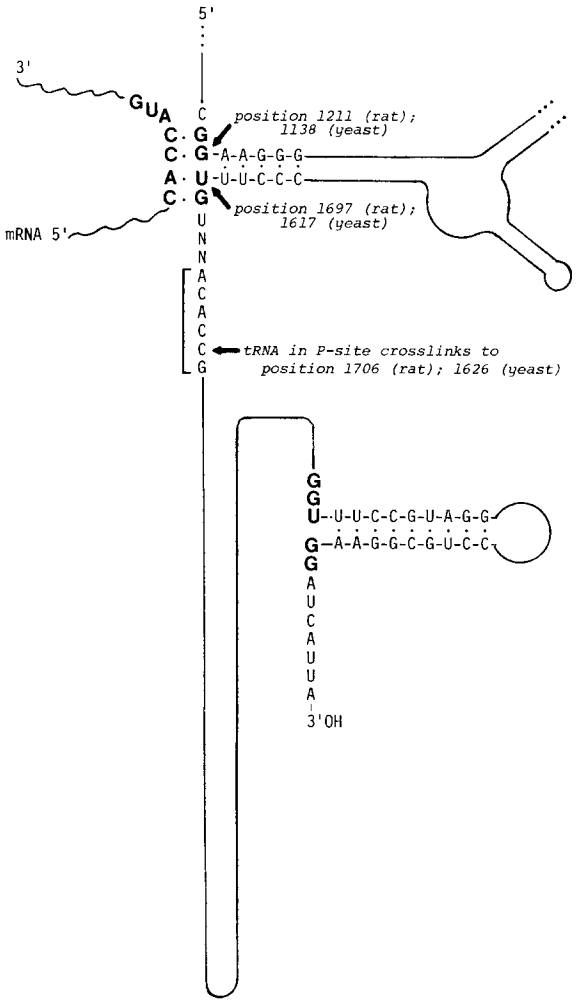

Figure 8. Possible Interaction between mRNA and 18S Ribosomal RNA

Highlighted are two sequences in 18S rRNA that are complementary to the conserved (CC)ACC motif in eukaryotic mRNAs. The GG/UGG sequence that extends across the base of the 3'-proximal hairpin in 18S rRNA was noted previously by Sargan et al. (1982). A novel interaction between mRNA and rRNA is shown close to the P site, where initiator Met-tRNA is believed to bind. The entire sequence of rat 18S rRNA, from which this segment was redrawn, is given in Chan et al. (1984).

quence in mRNA. Sargan et al. (1982) previously proposed that a noncontiguous GG/UGG sequence, brought together by the conserved hairpin near the 3' end of 18S rRNA, might pair with CCACC in mRNA. Their suggestion that the CCACC motif might function irrespective of its distance from the ATG codon now seems unlikely, but the site that they proposed in 18S rRNA remains interesting. The sequence 3'-GU/GG-5' occurs at the base of another hairpin in the interior of 18S rRNA, just a few nucleotides from the position to which peptidyl-tRNA becomes cross-linked when it is bound in the P site (Ehresmann et al., 1984). Initiator Met-tRNA also binds in the P site, of course. One could elaborate the model by suggesting that the nearby sequence ACACCG (bracketed in Figure 8), by virtue of being complementary to the proposed binding site for mRNA, might mediate a conformational switch within the rRNA that displaces the mRNA. In a similar vein, the region of E. coli 16S rRNA that includes the Shine-Dalgarno site is believed to undergo a rearrangement that ruptures the pairing with mRNA (Yuan et al., 1979). Nakashima et al. (1980) showed that eukaryotic mRNAs in 40S or 80S initiation complexes could be cross-linked by psoralen to 18S rRNA. Although they were specifically looking for an interaction between the 3' end of 18S rRNA and the cap-adjacent sequence in mRNA, their data actually fit better with an interaction between internal sequences in both RNAs.

The conservation of G in position +4 became evident several years ago, when the first few eukaryotic ribosome binding sites were sequenced. The suggestion (Kozak and Shatkin, 1977) that AUGG in mRNA might form a 4 base pair interaction with CCAU in the anticodon loop of initiator Met-tRNA has not yet been tested.

Experimental Procedures

Construction and Characterization of Mutants in the B and F Series

The parental plasmid p255/11 was derived previously (Kozak, 1984b) from Lomedico's original p255 (Lomedico and McAndrew, 1982) by deleting a Bam HI site at the junction between pBR322 and the rat genomic sequence. p255/11 retains a single Bam HI site, 8 bp down from the ATG initiator codon (Figure 1). After digesting p255/11 with Hind III and Bam HI, the large, linear fragment—which lacks only the four N-terminal amino acids of the preproinsulin coding sequence—was purified by agarose gel electrophoresis. This fragment was used as acceptor in DNA ligase reactions with various synthetic oligonucleotides, which were purchased from Pharmacia P-L Biochemicals. Each of the oligonucleotides listed in Table 1 was annealed with another oligonucleotide that was partly complementary, resulting in duplex structures with single-stranded termini complementary to the Hind III and Bam HI ends of the acceptor DNA. This is illustrated in Figure 1. In preparation for ligation, the oligonucleotides were preannealed but were not phosphorylated. The DNA ligase reaction typically contained 1 μg of linearized acceptor DNA and 0.2 A_{260} units of oligonucleotide in 25 μl. The mixture was incubated at 16°C for 20 hr, heated for 10 min at 65°C to inactivate the enzyme, and used directly to transform E. coli as described previously (Kozak, 1983b). DNA from ampicillin-resistant colonies was screened for the desired mutation by direct sequencing of plasmid DNA extracted rapidly from 10 ml cultures (Kozak, 1983b); the sequences were confirmed at a later step using pure DNA. For sequence analysis, DNA was labeled with α-^{32}P-CTP at the Dde I site that lies in the SV40 portion of the leader (Figure 1 in Kozak, 1984b), recut with Eco RI, and subjected to chemical cleavage (Maxam and Gilbert, 1980).

Construction of Reinitiation Derivatives

Several plasmids were constructed in which the ability to make proinsulin depends on reinitiating downstream from a terminator codon. The upstream "minicistron" (i.e., a fragment that contains an ATG initiator codon followed shortly by an in-phase terminator codon) was carried on a Hind III fragment from p255/21, which has been described previously (Kozak, 1984c). DNA from mutants B34, B35, B37 (a control), B38, and B39 was linearized by digesting with Hind III, treated with alkaline phosphatase, and the small 184 bp Hind III fragment from p255/21 was then inserted. This fragment includes a 119 bp intron; thus, the insert in mature mRNA is 65 nucleotides. The structures of the resulting "R" derivatives are shown in Figure 7. The number and orientation of inserts was determined by analysis with appropriate restriction enzymes.

Analysis of Proteins and RNA from Transfected COS Cells

COS-1 cells (Gluzman, 1981) in 60 mm plates were transfected one day after plating, when the monolayers were about 75% confluent. Each

plate was transfected with 0.5 ml of calcium phosphate mixture (Wigler et al., 1978) containing 20 μg of purified plasmid DNA and 40 μg of calf thymus carrier DNA. Other details of the procedure were as described previously (Kozak, 1983b). Forty-eight hours after transfection, the cells were rinsed twice with cysteine-free medium and incubated for 4 hr with 1 ml of medium containing 0.25 mCi of ^{35}S-cysteine (New England Nuclear). The cells were scraped with a rubber policeman into phosphate-buffered salts, and were then lysed with 0.9% NP40 and 0.4% deoxycholate. A 200 μl aliquot of the cytoplasmic extract, representing about 4 × 10^5 cells, was incubated at 4°C for 20 hr with 3 μl of antiserum against bovine insulin (Miles). The immunoprecipitates were recovered using Pansorbin (Calbiochem) and were analyzed by electrophoresis in polyacrylamide gels containing urea and SDS, as described (Kozak, 1983b). The gels were impregnated with Enhance (New England Nuclear), dried, and placed in contact with XAR-5 film for 2 to 20 days at −70°C. The data were quantified by densitometric scanning of the films.

Cytoplasmic RNA levels were analyzed by dot blot hybridizations using a ^{32}P-labeled riboprobe. The probe was obtained by cloning the 190 bp Bam HI–Eco RI fragment, which represents two-thirds of the coding sequence for rat preproinsulin, into the polylinker region of pSP65. The construct was linearized with Bam HI and transcribed with SP6 polymerase (New England BioLabs), according to Melton et al. (1984). The size of the transcript was checked by electrophoresis on 8% polyacrylamide gels containing 8 M urea. Cytoplasmic RNA from 48 hr transfected COS cells was extracted with phenol, serially diluted, and bound to Gene Screen filters (New England Nuclear). Prehybridization and hybridization were carried out using the conditions recommended by Melton et al. (1984).

Acknowledgments

I thank Jim Pipas for providing space in his cell culture facility. Research funds were provided by a grant from the National Institutes of Health (GM 33915).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *"advertisement"* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received August 28, 1985; revised October 28, 1985

References

Bandyopadhyay, P. K., and Temin, H. M. (1984). Expression from an internal AUG codon of herpes simplex thymidine kinase gene inserted in a retrovirus vector. Mol. Cell. Biol. *4*, 743–748.

Bellini, W. J., Englund, G., Rozenblatt, S., Arnheiter, H., and Richardson, C. D. (1985). Measles virus P gene codes for two proteins. J. Virol. *53*, 908–919.

Bos, J. L., Polder, L. J., Bernards, R., Schrier, P., van den Elsen, P., van der Eb., A., and van Ormondt, H. (1981). The 2.2 kb E1b mRNA of human Ad12 and Ad5 codes for two tumor antigens starting at different AUG triplets. Cell *27*, 121–131.

Castle, E., Nowak, T., Leidner, U., Wengler, G., and Wengler, G. (1985). Sequence analysis of the viral core protein and the membrane-associated proteins of flavivirus West Nile Virus. Virology, in press.

Chan, Y-L., Gutell, R., Noller, H. F., and Wool, I. G. (1984). The nucleotide sequence of a rat 18S ribosomal RNA gene and a proposal for the secondary structure of 18S rRNA. J. Biol. Chem. *259*, 224–230.

Clarke, B. E., Sangar, D. V., Burroughs, J. N., Newton, S. E., Carroll, A. R., and Rowlands, D. J. (1985). Two initiation sites for foot and mouth disease virus polyprotein in vivo. J. Gen. Virol. *145*, 227–236.

Clerx-van Haaster, C., Akashi, H., Auperin, D., and Bishop, D. (1982). Nucleotide sequence analyses and predicted coding of bunyavirus genome RNA species. J. Virol. *41*, 119–128.

Dixon, L., and Hohn, T. (1984). Initiation of translation of the cauliflower mosaic virus genome from a polycistronic mRNA: evidence from deletion mutagenesis. EMBO J. *3*, 2731–2736.

Ehresmann, C., Ehresmann, B., Millon, R., Ebel, J-P., Nurse, K., and Ofengand, J. (1984). Cross-linking of the anticodon of E. coli and B. subtilis acetylvalyl-tRNA to the ribosomal P site. Biochemistry *23*, 429–437.

Ernst, H., and Shatkin, A. J. (1985). Reovirus hemagglutinin mRNA codes for two polypeptides in overlapping reading frames. Proc. Natl. Acad. Sci. USA *82*, 48–52.

Giorgi, C., Blumberg, B., and Kolakofsky, D. (1983). Sendai virus contains overlapping genes expressed from a single mRNA. Cell *35*, 829–836.

Gluzman, Y. (1981). SV40-transformed simian cells support the replication of early SV40 mutants. Cell *23*, 175–182.

Heermann, K., Goldmann, U., Schwartz, W., Seyffarth, T., Baumgarten, H., and Gerlich, W. H. (1984). Large surface proteins of hepatitis B virus containing the pre-S sequence. J. Virol. *52*, 396–402.

Hughes, S., Mellstrom, K., Kosik, E., Tamanoi, F., and Brugge, J. (1984). Mutation of a termination codon affects *src* initiation. Mol. Cell. Biol. *4*, 1738–1746.

Konarska, M., Filipowicz, W., Domdey, H., and Gross, H. J. (1981). Binding of ribosomes to linear and circular forms of the 5′-terminal leader fragment of tobacco mosaic virus RNA. Eur. J. Biochem. *114*, 221–227.

Kozak, M. (1979). Inability of circular mRNA to attach to eukaryotic ribosomes. Nature *280*, 82–85.

Kozak, M. (1980a). Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. Cell *22*, 7–8.

Kozak, M. (1980b). Role of ATP in binding and migration of 40S ribosomal subunits. Cell *22*, 459–467.

Kozak, M. (1981). Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucl. Acids Res. *9*, 5233–5252.

Kozak, M. (1983a). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev. *47*, 1–45.

Kozak, M. (1983b). Translation of insulin-related polypeptides from mRNAs with tandemly reiterated copies of the ribosome binding site. Cell *34*, 971–978.

Kozak, M. (1984a). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucl. Acids Res. *12*, 857–872.

Kozak, M. (1984b). Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature *308*, 241–246.

Kozak, M. (1984c). Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. Nucl. Acids Res. *12*, 3873–3893.

Kozak, M. (1986). Influences of mRNA secondary structure on initiation by eucaryotic ribosomes. Proc. Natl. Acad. Sci. USA, in press.

Kozak, M., and Shatkin, A. J. (1977). Sequences and properties of two ribosome binding sites from the small size class of reovirus mRNA. J. Biol. Chem. *252*, 6895–6908.

Kozak, M., and Shatkin, A. J. (1978). Migration of 40S ribosomal subunits on mRNA in the presence of edeine. J. Biol. Chem. *253*, 6568–6577.

Krieg, P., Strachan, R., Wallis, E., Tabe, L., and Colman, A. (1984). Efficient expression of cloned complementary DNAs for secretory proteins after injection into Xenopus oocytes. J. Mol. Biol. *180*, 615–643.

Lamb, F. I., Roberts, L. M., and Lord, J. M. (1985). Nucleotide sequence of cloned cDNA coding for preproricin. Eur. J. Biochem. *148*, 265–270.

Laprevotte, I., Hampe, A., Sherr, C., and Galibert, F. (1984). Nucleotide sequence of the *gag* gene and *gag-pol* junction of feline leukemia virus. J. Virol. *50*, 884–894.

Liu, C-C., Simonsen, C. C., and Levinson, A. D. (1984). Initiation of translation at internal AUG codons in mammalian cells. Nature *309*, 82–85.

Lomedico, P., and McAndrew, S. (1982). Eukaryotic ribosomes can recognize preproinsulin initiation codons irrespective of their position relative to the 5′-end of mRNA. Nature *299*, 221–226.

Matteucci, M. D., and Heyneker, H. L. (1983). Targeted random mutagenesis: the use of ambiguously synthesized oligonucleotides to

mutagenize sequences immediately 5′ of an ATG initiation codon. Nucl. Acids Res. *11*, 3113–3121.

Maxam, A. M., and Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. Meth. Enzymol. *65*, 499–560.

Melton, D. A., Krieg, P. A., Rebagliati, M. R., Maniatis, T., Zinn, K., and Green, M. R. (1984). Efficient in vitro synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. Nucl. Acids Res. *12*, 7035–7056.

Morlé, F., Lopez, B., Henni, T., and Godet, J. (1985). α-Thalassaemia associated with the deletion of two nucleotides at position −2 and −3 preceding the AUG codon. EMBO J. *4*, 1245–1250.

Nakashima, K., Darzynkiewicz, E., and Shatkin, A. J. (1980). Proximity of mRNA 5′-region and 18S rRNA in eukaryotic initiation complexes. Nature *286*, 226–230.

Persing, D. H., Varmus, H. E., and Ganem, D. (1985). A frameshift mutation in the pre-S region of the human hepatitis B virus genome allows production of surface antigen particles but eliminates binding to polymerized albumin. Proc. Natl. Acad. Sci. USA *82*, 3440–3444.

Reddy, V. B., Dhar, R., and Weissman, S. M. (1978). Nucleotide sequence of the genes for the simian virus 40 proteins VP2 and VP3. J. Biol. Chem. *253*, 621–630.

Sargan, D. R., Gregory, S. P., and Butterworth, P. H. W. (1982). A possible novel interaction between the 3′-end of 18S rRNA and the 5′-leader sequence of many eukaryotic mRNAs. FEBS Letters *147*, 133–136.

Sarkar, G., Pelletier, J., Bassel-Duby, R., Jayasuriya, A., Fields, B. N., and Sonenberg, N. (1985). Identification of a new polypeptide coded by reovirus gene s1. J. Virol. *54*, 720–725.

Shatkin, A. J. (1976). Capping of eucaryotic mRNAs. Cell *9*, 645–653.

Smith, A. E. (1977). Cryptic initiation sites in eukaryotic virus mRNAs. Federation of European Biological Societies Symposium *43*, 37–46.

Smith, G. E., Summers, M. D., and Fraser, M. J. (1983). Production of human beta interferon in insect cells infected with a baculovirus expression vector. Molec. Cell. Biol. *3*, 2156–2165.

Wigler, M., Pellicer, A., Silverstein, S., and Axel, R. (1978). Biochemical transfer of single-copy eucaryotic genes using total cellular DNA as donor. Cell *14*, 725–731.

Yuan, R. C., Steitz, J. A., Moore, P. B., and Crothers, D. M. (1979). The 3′ terminus of 16S rRNA: secondary structure and interaction with ribosomal protein S1. Nucl. Acids Res. *7*, 2399–2418.