Strength of the Translation Initiation Signal Sequence of mRNA as Studied by the Quantification Method: Effect of Nucleotides at Positions from +4 to +6 upon the Recognition of ATG and Alternative Initiation Codons

Yôichi Iida* **and Daisuke Kanagu**

Division of Chemistry, Graduate School of Science, Hokkaido University, Sapporo 060-0810

(Received June 17, 2002)

Concerning the translation initiation signal in vertebrate mRNAs, although a consensus sequence, (GCC)GCC(A or G)CCATGG, has been proposed, the actual initiation sequences differ from it to a greater or lesser degree. Kozak monitored selection by ribosomes of the first versus second ATG codons as a function of introducing mutations at positions +4, +5 and +6 of the first ATG codon. Six different flanking codons, starting with G, strongly enhanced the selection of the first ATG codon. Except for T at position +5, ATG codon recognition was unaffected by most mutations at positions +5 and +6. These experimental results were analyzed using a quantification method proposed previously, and a 18-nucleotide sequence including ATG and nucleotides in positions +4, +5 and +6 was characterized by its sample score (strength of translation initiation signal). The experimental results were well understood based on the strength of the signal.

In the process of translating eukaryotic mRNAs, the 40S ribosomal subunits appear to bind first at the 5′-end (cap site) of mRNA and to scan the mRNA until the subunits find an AUG translation initiation codon (hereafter, u or U of mRNA is often described by t or T of cDNA, respectively). Although it has been generally believed that translation initiation occurs at the ATG codon nearest to the cap site, the downstream second ATG often functions. In addition to the invariant ATG codon, a considerable sequence homology is found around the 5′-untranslated region flanking the ATG codon. Kozak[1] proposed a consensus sequence, (GCC)GCC(A or G)CCATGG, as the optimal sequence for the translation initiation by vertebrate ribosomes, where the underlined ATG (hereafter, the position of A is denoted by +1) is the invariant initiation codon. Sequence flanking the ATG modulates its ability to halt the scanning 40S ribosomal subunit. Mutations that particularly weaken the 40S subunit to bind the consensus sequence are substitutions of a nucleotide of A at position −3 and G at position +4. These substitutions cause some 40S ribosomal subunits to bypass the first ATG and to initiate instead at the next downstream ATG. However, recent studies suggested that translation initiation might be affected by nucleotides extending into the coding domain beyond +4. For example, Grunert and Jackson[2] reported that initiation at an ATG or a CTG start codon was favored by A at position +5 and T at +6. It is very difficult to examine the involvement of these nucleotides, because mutations introduced into such positions of the mRNA may change the amino acid sequence, thus changing the stability of the encoded protein. To avoid this problem, Kozak[3] developed an assay which directly monitors ribosome-mRNA complexes, and obtained experimental data on the effect of nucleotide changes at positions +4, +5 and +6 upon the efficiency of recognition of the ATG initiation codon.

On the other hand, we previously analyzed the initiation signal sequences of rat preproinsulin mRNAs by a quantification method, and proposed strength of the signal.[4] At that time, however, we only examined 16-nucleotide sequence including ATG and nucleotides at positions from −12 to +4. The calculated strength of the signal well explained the experimental translation initiation efficiency. In the present work, we applied such an approach to Kozak's data on the effect of nucleotide changes at positions +4, +5 and +6. An excellent agreement was obtained between her experimental translation efficiency and our strength of the signal.

Results and Discussion

Kozak's Experiments on Mutations at Positions +4, +5 and +6. Kozak developed a primer extension assay to monitor selection by ribosomes of the first versus the second ATG codon as a function of introducing mutations at positions +4, +5 and +6.[3] The mRNA used for the experiment has two initiation codons and two open reading frames (ORFs), as outlined in Fig. 1A. ORF1 extends from ATG#1 to a TAA stop codon overlapping Lue45 in the chloramphenicol acetyltransferase (CAT) coding sequence, and encodes a 70-amino-acid polypeptide with a molecular weight of 8 kDa. Its product is out-of-frame with respect to CAT, and is designated by p8. ORF2 initiates with ATG#2, and is in-frame with the downstream CAT coding sequence. ORF2 encodes a 240-amino-acid polypeptide with a molecular weight of 28 kDa, and its product is designated by p28. She incubated such a mRNA in a rabbit reticulocyte translation system with inhibitors to block elongation. Ribosome-mRNA binding at the ATG initiation codon can be monitored directly by the amount of initiation

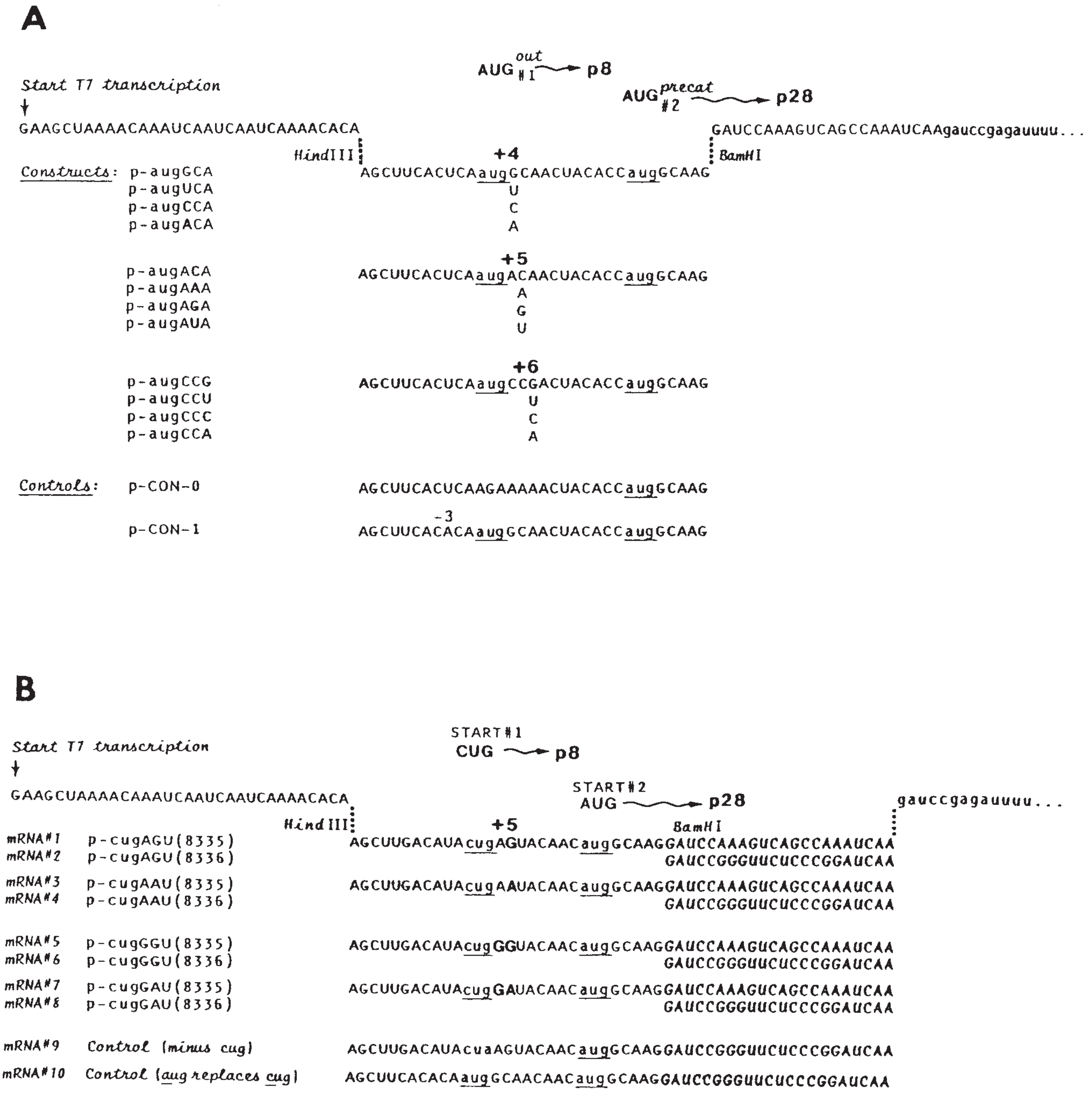
A
Start T7 transcription
GAAGCUAAAACAAAUCAAUCAAUCAAAACACA
AUG #1 out p8
AUG #2 precat p28
GAUCCAAAGUCAGCCAAAUCAAgauccgagauuuu...
HindIII
BamHI
+4
Constructs: p-augGCA
p-augUCA
p-augCCA
p-augACA
AGCUUCACUCAaugGCAACUACACCaugGCAAG
U
C
A
+5
p-augACA
p-augAAA
p-augAGA
p-augAUA
AGCUUCACUCAaugACAACUACACCaugGCAAG
A
G
U
+6
p-augCCG
p-augCCU
p-augCCC
p-augCCA
AGCUUCACUCAaugCCGACUACACCaugGCAAG
U
C
A
Controls: p-CON-0
AGCUUCACUCAAGAAAAACUACACCaugGCAAG
-3
p-CON-1
AGCUUCACACAaugGCAACUACACCaugGCAAG
B
Start T7 transcription
GAAGCUAAAACAAAUCAAUCAAUCAAAACACA
START#1 CUG p8
START#2 AUG p28
gauccgagauuuu...
HindIII
+5
BamHI
mRNA#1 p-cugAGU(8335)
mRNA#2 p-cugAGU(8336)
AGCUUGACAUAcugAGUACAACaugGCAAGGAUCCAAAGUCAGCCAAAUCAA
GAUCCGGGUUCUCCCGGAUCAA
mRNA#3 p-cugAAU(8335)
mRNA#4 p-cugAAU(8336)
AGCUUGACAUAcugAAUACAACaugGCAAGGAUCCAAAGUCAGCCAAAUCAA
GAUCCGGGUUCUCCCGGAUCAA
mRNA#5 p-cugGGU(8335)
mRNA#6 p-cugGGU(8336)
AGCUUGACAUAcugGGUACAACaugGCAAGGAUCCAAAGUCAGCCAAAUCAA
GAUCCGGGUUCUCCCGGAUCAA
mRNA#7 p-cugGAU(8335)
mRNA#8 p-cugGAU(8336)
AGCUUGACAUAcugGAUACAACaugGCAAGGAUCCAAAGUCAGCCAAAUCAA
GAUCCGGGUUCUCCCGGAUCAA
mRNA#9 Control (minus cug)
AGCUUGACAUAcuaAGUACAACaugGCAAGGAUCCGGGUUCUCCCGGAUCAA
mRNA#10 Control (aug replaces cug)
AGCUUCACACAaugGCAACAACaugGCAAGGAUCCGGGUUCUCCCGGAUCAA

Fig. 1. (A) Sequences of mRNAs used to study effects of varying nucleotides in positions +4, +5 and +6. The 3′-end of the sequences lead to the CAT coding domain. Mutations were introduced around the first AUG codon, which initiate translation of an 8 kDa polypeptide (p8). The second AUG codon initiates translation of a 28 kDa polypeptide (p28). (B) Sequences of mRNAs used for translation at a CUG initiation codon. These figures are taken from Figs. 1 and 7 of Kozak.[3] Refer to their captions for details of the figures. In our report, u or U of mRNA is described by t or T of cDNA, respectively.

complexes accumulated at the ATG codon. The particular ATG start site is identified by using a primer extension inhibition assay in which a ^{32}P-labeled deoxyoligonucleotide primer, annealed to the mRNA downstream from all potential initiator codons, is extended by reverse transcriptase up to the 3′ edge of the bound ribosome. The strength of ATG start codon recognition can be measured in terms of the amount of primer extension product. Kozak monitored the ratio of initiation at ATG#1/ATG#2 as a function of introducing mutations around ATG#1; the experimental data will be discussed in the following section.

Quantification Analysis of Translation Initiation Signal Sequences. An earlier quantification analysis was conducted to study the translation initiation signal sequences of rat preproinsulin and its mutant mRNAs.[4] For this purpose, the previous paper discussed the 16-nucleotide sequence from positions −12 to +4, which is composed of 12 nucleotides in the 5′-untranslated region, 3 in the ATG initiation codon and

Table 1. The 18-Nucleotide Sequence Data of Translation Initiation Signal in Vertebrate mRNAs to be Analyzed by Quantification Method

No. (ν)	Group (r)[a)]	Sequence	Gene
1	1	CTCCGGTAGCCCATGGAG	Human acetylchloline receptor
2	1	CGGGTGTTAGACATGGGT	Rat acetylchloline receptor
:	:	:	
698	1	CATTTGGAAAGGATGTAC	Xenopus pre-xenopsin
699	1	AAGCTTCACTCAATGGCA	p-atgGCA ATG#1
700	1	GGCAACTACACCATGGCA	p-atgGCA ATG#2
1	2	GAAGCTAAAACAAATCAA	p-atgGCA
2	2	AAGCTAAAACAAATCAAT	
:	:	:	
851	2	ATGGCAGAAATTCGGATC	

a) Group 1 is composed of authentic translation initiation signal sequences in vertebrate mRNAs, while group 2 comprises sequences including no such signal. Sequences of group 1 are taken from Kozak.[1] Sequences of group 2 are constructed by using p-atgGCA mRNA. See Fig. 1A and text for further details.

1 in the coding region. A set of 700 signal sequences was used, which were taken from sequences at the authentic initiation sites in various vertebrate mRNAs, as compiled by Kozak.[1] To analyze the effect of nucleotides at positions +4, +5 and +6, a new set of 700 signal sequences of 18-nucleotides from positions −12 to +6 is necessary. We compiled such a data set by adding sequences between +4 and +6 to Kozak's data.

To analyze the translation initiation signal by a quantification method, we constructed two groups of sequence data, as given in Table 1. The first group ($r = 1$) is composed of 700 samples of the preceding 18-nucleotide sequences, which include a translation initiation signal. The sequences in the second group ($r = 2$) do not include such a signal. They were taken from the mRNA sequence in which Kozak studied the effects of varying the nucleotide at positions +4, +5 and +6 (see Fig. 1A).[3] For example, in the mRNA sequence of a construct of p-atgGCA, we start with the 18-nucleotide sequence at the cap site. Next, we progress one nucleotide in the 3′-direction, and take the next 18-nucleotide sequence. In this way, we window 18-nucleotide sequence at every position down to the 3′-end of the CAT sequence. In such a sequence, however, two sequences lie at the authentic translation initiation sites of ATG#1 and ATG#2. They are brought into the first group ($r = 1$; $\nu = 699$ and 700, respectively). The remaining 851 sample sequences are summarized into the second group (see Table 1). Next, the sequence data are transformed into item-category data. For this purpose, we introduce a dummy variable, $x^{r(\nu)}_{i(\alpha)}$, which is defined by item ($i = 1, 2, \ldots, 18$), category ($\alpha = 1, 2, 3, 4$), group ($r = 1, 2$), and sample ($\nu = 1, 2, \ldots, n_r$). Eighteen items correspond to the positions of nucleotides in the 18-nucleotide sequences, i being given by the order from the 5′- to 3′-ends of the sequence. Four categories denote the kinds of nucleotides, where A, G, C or T is specified by $\alpha = 1, 2, 3$ or 4 at every item, respectively. Parameter, ν, specifies each sample sequence belonging to the group ($r = 1$ or 2); $n_1 = 700$ and $n_2 = 851$ are total number of sample sequences in each group. The dummy variable, $x^{r(\nu)}_{i(\alpha)}$, takes 1, if the sample sequence (ν) of the group (r) has a nucleotide (α) at position (i); otherwise it takes 0. Using this variable, we transform the sequence data of Table 1 into item-category data composed of 0 or 1.

Quantification of each sequence can be done by calculating the sample score value,

$$y^{r(\nu)} = \sum_{i=1}^{18} \sum_{\alpha=1}^{4} x^{r(\nu)}_{i(\alpha)}\, a_{i(\alpha)}, \tag{1}$$

where $r = 1, 2$ and $\nu = 1, 2, \ldots, n_r$. The coefficient of $a_{i(\alpha)}$ is a real number and is called the category weight. Our quantification method determines the $a_{i(\alpha)}$ and $y^{r(\nu)}$ values in such a way that the two groups of sequences including translation initiation signal ($r = 1$) and sequences including no such signal ($r = 2$) may be discriminated most distinctly. This optimization can be achieved by the following procedure. First, we calculate the mean value of sample scores within the group r, $\bar{y}^r$, and the mean value of the total samples, $\bar{y}$. Then, the variance of the total samples, σ^2, and the variance between groups 1 and 2, σ^2_{B}, are given by

$$\sigma^2 = (1/N) \sum_{r=1}^{2} \sum_{\nu=1}^{n_r} (y^{r(\nu)} - \bar{y})^2, \tag{2}$$

$$\sigma^2_{\mathrm{B}} = (1/N) \sum_{r=1}^{2} n_r (\bar{y}^r - \bar{y})^2, \tag{3}$$

where $N = n_1 + n_2$. To discriminate the sequences between groups 1 and 2 most distinctly, we maximize the $\sigma^2_{\mathrm{B}}/\sigma^2$ value. Estimation of $a_{i(\alpha)}$ values at this optimum condition can be done by solving the eigen-value problem, and the procedure was described previously.[5] Values of $a_{i(\alpha)}$ thus estimated for the p-atgGCA mRNA are given in Table 2. Sample score of any 18-nucleotide sequence in the mRNA is then calculated by Eq. 1 together with the $a_{i(\alpha)}$ values. Our analysis demonstrates that the greater is the score of a sequence, the stronger is the translation initiation signal of the sequence. Using the same approach, we analyzed 18-nucleotide sequences of Kozak's other mutant mRNAs.

Quantification Analysis of Experimental Data of mRNAs with Mutations in Positions +4, +5 and +6. Using an assay which directly monitors the initiation step of translation, Kozak examined the effect on the recognition of the ATG#1

Table 2. Optimum Category Weight Values of $a_{i(\alpha)}$ for Translation Initiation Signal Calculated by Quantification Analysis of the Data of Table 1[a)]

Item (*i*)	Category (α)/nucleotide			
	1/A	2/G	3/C	4/T
1	−0.2721	0.0969	0.3863	−0.2408
2	−0.1189	0.2006	0.2815	−0.3892
3	−0.0314	0.0365	0.1814	−0.2142
4	−0.4493	0.2415	0.3397	−0.1208
5	−0.2540	0.0014	0.3873	−0.2253
6	0.1786	−0.2533	0.3319	−0.3752
7	−0.4504	0.1982	0.1870	0.0194
8	−0.2204	−0.0659	0.3451	−0.1424
9	−0.0137	−0.2128	0.4122	−0.5720
10	0.6979	0.2052	−1.2689	−1.2803
11	0.2884	−0.2647	0.4098	−0.8726
12	−0.4300	0.0084	0.5095	−0.4857
13	1.3878	−2.1918	−2.1655	−2.1100
14	−2.3601	−2.3581	−2.1764	1.5348
15	−2.5262	2.0167	−2.5171	−3.0701
16	−0.2284	0.6073	−0.4484	−0.2619
17	0.0677	0.0788	0.3406	−0.6481
18	−0.4106	0.3510	−0.0749	0.0127

a) Item number (*i*) specifies the position of nucleotide, while category number (α), the kind of nucleotide. For further details, see text.

Table 3. Comparison between Experimental ATG#1/ATG#2 Ratio and Sample Score Calculated from Quantification Method in a Variety of Codons Flanking ATG#1

Mutant[a)]	ATG#1/ATG#2[b)]	Sample score
p-atgGCA	2.6	3.9396
T	1.0	3.1239
C	0.8	2.9185
A	0.9	3.1689
p-atgACA	1.1	3.1689
A	1.0	2.8795
G	1.0	2.9592
T	1.0	2.2378
p-atgCCG	1.1	3.6525
T	1.0	3.2862
C	0.9	3.2085
A	0.8	2.9185

a) See Fig. 1A. b) Ratio of initiation at ATG#1 versus ATG#2. Values were normalized to the mRNA in each series that has T in the test position. Because the sequence flanking ATG#2 was constant, an increase in the ATG#1/ATG#2 ratio indicates improved recognition of ATG#1. For details of experimental condition, see Table 1 in Kozak.[3]

codon when positions +4, +5 and +6 were varied systematically.[3] As is shown in Table 3, the codon following ATG#1 was NCA for the +4 series, ANA for the +5 series and CCN for the +6 series, where N is A, G, C or T. The mRNA sequences are fully given in Fig. 1A. Because the sequence flanking ATG#2 was constant, an increase in ATG#1/ATG#2 indicates improved recognition of ATG#1. The experimental data show a strongly positive effect of G^{+4} in the series ACA, GCA, CCA and TCA, but no effect on recognition was found when position +5 was found in the series ACA, AAA, AGA and ATA. The efficiency of ATG#1 was also affected very little when position +6 was varied in the series CCG, CCT, CCC and CCA.

These experimental results can be explained by our quantification analysis. In Table 3, we compare Kozak's experimental data with the calculated sample scores of 18-nucleotide sequences in the series of constructs given in Fig. 1A. For example, in the p-atgGCA construct, the 18-nucleotide sequence including ATG#1 (underlined) is given by CAGCTTCACTCAATGGCA. The score of this sequence was calculated to be 3.9396 by using the category weight data of Table 2. In the other constructs given in Fig. 1A, we also calculated the sample scores of the 18-nucleotide sequences including ATG#1; their values are summarized in Table 3. The calculated sample scores were then compared with the translation efficiencies of four mRNAs that differ in a single position downstream of ATG#1 at a given concentration of 2.0 mM Mg^{2+}. In the p-atgNCA series, ratio of ATG#1/ATG#2 has the greatest value of 2.6 with N = G. In accordance with this, the score has the greatest value of 3.9396. The ATG#1/ATG#2 values decrease to 1.0 (N = T), 0.9 (N = A) and 0.8 (N = C); almost in parallel with this, the sample scores decrease to 3.1293, 3.1689 and 2.9185, respectively. In the p-atgANA series, the ATG#1/ATG#2 = 1.1 value is greatest with N = C, while it is 1.0 with N = A, G or T. In accordance with this, the score of 3.1689 is the greatest with N = C, while score is 2.8795, 2.9592 or 2.2378 with N = A, G or T, respectively. In the p-atgCCN series, ATG#1/ATG#2 = 1.1 is the greatest with N = G, and decreases progressively to 1.0, 0.9 and 0.8 with N = T, C and A, respectively. The score 3.6525 is the greatest with N = G, and decreases progressively to 3.2862, 3.2085 and 2.9185 with N = T, C and A, respectively. The order of the decrease in ATG#1/ATG#2 agrees well with that of the sample scores. All of these results indicate that G in position +4 gives both exceptionally great ATG#1/ATG#2 and sample score values. Although nucleotide substitutions at positions +5 and +6 affect both ATG#1/ATG#2 and sample score values, the effect of substitutions is not as important as that of +4 substitutions, and the optimal context for initiation does not extend beyond G in position +4.

Although above experiments show a strong positive effect of G^{+4}, only p-atgGCA was examined, and there remains a possibility that the flanking codon GCA may happen to favor initiation. To exclude this possibility, Kozak examined six different GNN codons adjacent to ATG#1.[3] Her experimental ATG#1/ATG#2 values are summarized in Table 4, where each mRNA is compared with a matched construct that had C or A instead of G in +4 position, and where the actual values of ATG#1/ATG#2 are given by autoradiogram experiments. Table 4 clearly shows that ATG#1 was recognized by about 3-fold better in five out of six cases, where G^{+4} was the flanking nucleotide. Since five different flanking codons starting with G(GCG, GCT, GCC, GCA and GAT) strongly enhanced the recognition of ATG#1, the enhancement was attributable to G^{+4} rather than to a particular flanking codon. These experimental results were also explained by our quantification analysis. As shown in Table 4, a calculation of sample scores revealed that the 18-nucleotide sequences containing GCG,

Table 4. Comparison between Experimental ATG#1/ATG#2 Ratio and Sample Score to Show Positive Effect of G in Position +4 in a Variety of Codons Flanking ATG#1

Mutant[a)]	ATG#1/ATG#2[b)]	Sample score
p-atgGCG	2.00	4.6797
C	0.61	3.6525
p-atgGCT	1.70	4.3121
C	0.41	3.2862
p-atgGCC	1.60	4.2458
C	0.47	3.2085
p-atgGCA	1.40	3.9396
C	0.46	2.9185
p-atgGAT	1.20	4.0394
A	0.50	3.2550
p-atgGTA	0.35	3.0077
A	0.40	2.2378

a) See Fig. 1A and text. b) Actual value for the ratio of initiation at ATG#1 versus ATG#2. For details of experimental condition, see Table 2 in Kozak.[3]

GCT, GCC, GCA and GAT flanking codons possess much larger scores than those containing CCG, CCT, CCC, CCA and AAT codons, respectively. For example, the 18-nucleotide sequence with GAT codon has a score of 4.0394, while that of AAT is 3.2550. The ATG#1/ATG#2 values of p-atgGAT and p-atgAAT are 1.20 and 0.50, respectively. A strong enhancement of G^{+4} was evident by both of the sample scores and the ATG#1/ATG#2 values.

In Table 4, we compare the ATG#1/ATG#2 values of six different mutants possessing G^{+4}, and find that the values range in the order of 2.00 (p-atgGCG) > 1.70 (p-atgGCT) > 1.60 (p-atgGCC) > 1.40 (p-atgGCA) > 1.20 (p-atgGAT) > 0.35 (p-atgGTA). It is very interesting to note that this order corresponds well to the order of sample score values, 4.6797 (p-atgGCG) > 4.3121 (p-atgGCT) > 4.2458 (p-atgGCC) > 4.0394 (p-atgGAT) > 3.9396 (p-atgGCA) > 3.0077 (p-atgGTA), except that the order of p-atgGCA and p-atgGAT is reversed. Kozak's experimental results are again explained in terms of our sample scores.

We notice that p-atgGTA was the only mRNA in which G^{+4} unexpectedly failed to enhance initiation. To determine if the flanking codon GTA specifically disfavors initiation, Kozak tested initiation at p-atgGTG, p-atgGTT and p-atgGTC together with p-atgGTA, and found that all four constructs show equally poor recognition of ATG#1.[3] In spite of the usual stimulatory effect of G^{+4}, the efficiency of ATG#1 initiation decreases greatly if G^{+4} is followed by T^{5+}. Fig. 6B of Kozak shows the experimental results of primer extension analysis of GAT, GTG, GTT, GTC, GTA and GGA codons following a ATG#1, where products due to the ATG#1 codon are very abundant with GAT and GGA codons. However, only small amounts of products are found with GTG, GTT, GTC and GTA codons. These experimental results are also explained by our quantification analysis. We calculated sample scores of 18-nucleotide sequences of those constructs, and compared scores with relative amounts of products, ATG#1/ATG#2. Both sample score and ATG#1/ATG#2 are the greatest with GAT codon, and the next greatest are the GGA codon. In contrast to these, scores of 3.3362 (GTT), 3.3188 (GTC) and 3.0077 (GTA) are smaller than 4.0394 (GAT) and 3.7074 (GGA). This is in good accordance with the finding that ATG#1/ATG#2 values of GTT, GTC and GTA are also very small in comparison with those of GAT and GGA. One exception is that score of 3.7379 of GTG is a little larger than 3.7074 of GGA, while ATG#1/ATG#2 of GTG is much smaller than that of GGA. Although this exception is not explained by our quantification analysis, both a primer extension analysis and a quantification analysis tell us that atgGT is not a favorable context for initiation. Kozak interpreted this as meaning that sequence GT at positions +4 and +5 may distort the mRNA, and thus impair ATG codon recognition by the scanning 40S ribosomal subunit.

Analysis of Non-ATG Start Sites. Instead of ATG, translation initiation sometimes occurs at CTG and GTG codons. In Kozak's experiments,[3] initiation at a CTG codon was detectable in the presence of G^{+4}. Being similar to the ATG start site, the strong dependence on the downstream sequence was also found at the CTG start site. Although poor initiation at CTG codon makes quantitative analysis difficult, experiments on the CTG start site confirmed a conclusion similar to that of the ATG start site. In the following, experimental results of CTG initiation codon were also analyzed by our quantification analysis.

The mRNAs used for such experiments are reproduced in Fig. 1B, where CTG replaces ATG#1 as the start codon for p8 polypeptide product. When these mRNAs were used as templates in a standard translation assay, production of some [^{3}H] leucine-labeled p8 was observed by the autoradiogram. Although the CTG start codon is much weaker than the ATG start codon, the CTG codon still inhibits downstream ATG#2 start codon and p28 synthesis. Because approximately positioned downstream secondary structure aids the recognition of weak CTG start codon, Kozak tested the CTG mRNAs with both an unstructured sequence (oligonucleotide-8335) and a structure-forming sequence (oligonucleotide-8336), and that the latter sequence significantly elevated initiation from the CTG codon. She prepared two other constructs for a comparison. In one construct, p-ctaAGT, where the CTG codon was mutated to CTA, p8 was completely absent, thus confirming that CTG is the source of p8. Another construct of p-atgGCA has the usual ATG codon. In this construct, only p8 was produced, and p28 was absent completely, thus showing that ATG is a strong start codon. In the following, we examine the data of the oligonucleotide-8336 sequence given in Figs. 7B and 7C of Kozak.[3] In view of the relative productions of p8 and p28, G in position +4 remarkably enhanced the recognition of the upstream CTG codon (compare ctgAGT with ctgGGT or compare ctgAAT with ctgGAT), whereas there was no appreciable difference between G and A at position +5 (compare ctgAGT with ctgAAT or compare ctgGGT with ctgGAT). Her study did not support the results of Grunert and Jackson,[2] who reported that the CTG start codon was favored by A at position +5. Our quantification analysis supports Kozak's experimental results. Sample scores of 18-nucleotide sequences were calculated with mRNAs of p-ctgAGT (0.5711), p-ctgAAT (0.5009), p-ctgGGT (1.3220), p-ctgGAT (1.2914) together with mRNA lacking ctg start codon, p-ctaAGT (−4.2296), and mRNA having usual ATG codon, p-atgGCA (5.9266). The score of the 18-nucleotide sequence lacking ctg is so small (−4.2296) that signal of initiation may be completely lost in such a mRNA. This is in good agreement with the finding for the mRNA, where p8 was completely absent, and where p28 was the only product arising from the ATG#2 start codon. On the other hand, the score of p-atgGCA (5.9266) was much greater than those of mRNAs having the CTG start codon. This is in good agreement with finding that, in p-atgGCA, p8 was the only product arising from ATG#1, and p28 was completely absent. Among the mRNAs having the CTG start codon, scores of 0.5009 of p-ctgAAT and 0.5711 of p-ctgAGT are the smallest, while 1.3220 of p-ctgGGT is the largest, and 1.2914 of p-ctgGAT, the next largest. These score values indicate that G at position +4 remarkably enhances the recognition of the CTG initiation codon, and that the score decreases if G is substituted for A at position +4. The results of a quantification analysis also agree well with Kozak's experimental results, mentioned above. As for the nucleotides at position +5, the scores of p-ctgAAT and p-ctgGAT are slightly smaller than those of p-ctgAGT and p-ctgGGT, respectively, and the substitution of A by G may improve the recognition of the initiator codon slightly. This does not agree with the finding that autoradiograms of p-ctgAAT and p-ctgGAT are a little more than those of p-ctgAGT and p-ctgGGT, respectively. However, such a difference is not a discernible one, and both the sample score and autoradiogram data show no positive effect of A at position +5, contradicting the hypothesis that initiation at the non-ATG codon might be favored by A instead of G at position +5.

Concluding Remarks

Selection by ribosomes of the first versus the second ATG codon was studied as a function of introducing mutations on the 3′ side sequence of the first ATG codon. Both a primer extension assay and a quantification analysis strongly support enhanced selection of the first ATG codon if mRNAs have G instead of A or C at position +4. In contrast to the enhancing effect of G, most mutations at positions +5 and +6 showed no appreciable effect. However, the enhancing effect of G at position +4 failed only if it was combined with T at position +5. These data were well explained by the strength of the translation initiation signal (score value of 18-nucleotide signal sequence). The strong positive effect of G at position +4 and a strong negative effect of T at position +5 are interesting, because no appreciable effect was found in other nucleotides at positions from +4 to +6. Kozak's experiment monitored directly ribosome-RNA complex,[3] and sequence GT at positions +4 and +5 may distort the mRNA. The sequence regions from −12 to −1 and from +4 to +6 appear to be recognized by protein-RNA interaction, and a certain protein included in ribosome has a tendency to recognize G^{+4} positively and T^{+5} negatively. Thus, GT at positions +4 and +5 impairs the ATG codon recognition by the scanning 40S ribosomal subunit.

In addition to the sequence context which increases the recognition of the ATG start codon, an enhancing effect of G at position +4 and no appreciable effect of nucleotides at position +5 were again observed with the CTG start codon. These observations were also supported by our quantification analysis.

Finally, we discuss the present results and those obtained in a previous paper of Iida and Masuda,[4] where they analyzed the translation initiation signal by using the 16-nucleotide sequence data of the first group ($r = 1$) taken from Kozak.[1] In the present work, we constructed 18-nucleotide sequence data of the first group by adding two nucleotides at positions of +5 and +6 to the same data set of Kozak. Since the lengths of sequences are different between the previous and present papers, we cannot directly compare the magnitudes of the category weight values, $a_{i(\alpha)}$, between them. However, the relative values of $a_{i(\alpha)}$ at positions from −12 to +4 in Iida and Masuda correspond well with those in the present work, because the sequence data of the first group in such a region are common.

References

1 M. Kozak, *Nucleic Acids Res.*, **15**, 8125 (1987).

2 S. Grunert and R. J. Jackson, *EMBO J.*, **13**, 3618 (1994).

3 M. Kozak, *EMBO J.*, **16**, 2482 (1997).

4 Y. Iida and T. Masuda, *Nucleic Acids Res.*, **24**, 3313 (1996).

5 Y. Iida, *Comput. Appl. Biosci.*, **3**, 93 (1987).