Nucleic Acids Research, 1996, Vol. 24, No. 17 **3313–3316**

Strength of translation initiation signal sequence of mRNA as studied by quantification method: effect of nucleotide substitutions upon translation efficiency in rat preproinsulin mRNA

Yôichi Iida* and Takeshi Masuda

Department of Chemistry, Graduate School of Science, Hokkaido University, Sapporo 060, Japan

Received June 3, 1996; Accepted June 14, 1996

ABSTRACT

Concerning the translation initiation signals in vertebrate mRNAs, both the ATG initiation codon and the sequences flanking the initiation codon are required to direct the position of initiation. A consensus sequence for the signal, (GCC)GCC(A or G)CCATGG, has been proposed, but actual initiation sequences differ from it to a greater or lesser degree. In the present report, the translation initiation signal sequences of rat preproinsulin and its mutant mRNAs were analyzed using a quantification method proposed previously. In this method, each 16 nt sequence in the mRNA was characterized by its sample score, which shows strength of the signal. So far, Kozak has constructed a number of preproinsulin mutant mRNAs in which nucleotides flanking the ATG codon are systematically varied, and measured the translation initiation efficiency in terms of the proinsulin product. Her experimental results were well understood on the basis of the strength of the translation initiation signal sequence.

INTRODUCTION

In the process of translation of 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 is described by T). Although it has been generally believed that the translation initiation occurs at the ATG codon nearest 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 is the invariant initiation codon. In the whole mRNA, however, there lie a number of sequences which resemble the consensus sequence; a problem arises, therefore, as to which sequence is chosen as the functional initiation signal. In the present report, we analyzed the initiation signal sequences of vertebrate mRNAs by the quantification method, and proposed strength of the signal. We applied such an approach to rat preproinsulin and its mutant mRNAs, where Kozak introduced systematic base substitutions into the sequence surrounding the initiation codon, and examined the translation initiation efficiency experimentally (2–4). Our quantification analysis shows an excellent correlation between her experimental translation efficiency and our strength of the signal.

QUANTIFICATION ANALYSIS OF TRANSLATION INITIATION SIGNAL SEQUENCES

Quantification analysis was used to study signals for splicing reactions of mRNA precursors (5,6). We again applied it to the analysis of translation initiation signal sequences. Taking rat preproinsulin mRNA, we show the calculation procedure briefly. As indicated in the previous section, Kozak proposed the 13 nt consensus sequence. However, a problem remains if a longer sequence is required for the translation initiation signal. In our calculation, referring to the consensus sequence, we examined a 16 nt sequence composed of 12 nt in the 5′-untranslated region, 3 nt in the ATG initiation codon and 1 nt in the coding region. As shown in Table 1, we constructed two groups of the sequence data. The first group (r = 1) is composed of 699 sets of 16 nt sequences, which include a translation initiation signal. They were taken from sequences at the authentic initiation sites in various vertebrate mRNAs, as compiled by Kozak (1). Sequences in the second group (r = 2) do not include such a signal. They were taken from rat preproinsulin mRNA in the following way. First, we start with the 16 nt sequence at the cap site. Next, we progress 1 nt in the 3′-direction, and take the next 16 nt sequence. In this way, we window 16 nt sequences at every position of the whole mRNA. In those sequences, however, one sequence lies at the authentic translation initiation site, which is brought into the first group. The remaining 443 sequences are summarized in the second group (see Table 1).

* To whom correspondence should be addressed

Table 1. The 16 nt sequence data of translation initiation signal in vertebrate mRNAs to be analyzed by quantification method

No. (ν)	Group (r)[a]	Sequence	Gene
1	1	CTCCGGTAGCCCATGG	Human α-Nicotinic
:	:	:	:
88	1	TTCGCTCGCACCATGG	Rat Calmodulin
:	:	:	:
347	1	CCTACCATCACCATGG	Salmon Insulin
:	:	:	:
699	1	AAAGCTTGGTTTATGT	Rat Preproinsulin (B38 Mutant)
1	2	GGCCTCTGAGCTATTC	Rat Preproinsulin (B38 Mutant)
2	2	GCCTCTGAGCTATTCC	
3	2	CCTCTGAGCTATTCCA	
:	:	:	
443	2	AACCTTTGAAAGAGCA	

[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 the group 1 are taken from ref. 1. Sequences of the group 2 are constructed by using rat preproinsulin B38 mutant mRNA. See text for further details.

Next, the sequence data are transformed into item–category data. For this purpose, we introduce a dummy variable, $x_{i(\alpha)}^{r(\nu)}$, which is defined by item (i = 1, 2, ..., 16), category (α = 1, 2, 3, 4), group (r = 1, 2) and sample (ν = 1, 2, ..., n_r). Sixteen items correspond to the positions of nucleotides in the 16 nt 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 α = 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 = 699 and n_2 = 443 are the total number of sample sequences in each group. The dummy variable, $x_{i(\alpha)}^{r(\nu)}$, takes 1, if the sample sequence (ν) of the group (r) has a nucleotide (α) at the position (i); otherwise it takes 0. Using this variable, we transform the sequence data of Table 1 into the item–category data composed of 0 or 1.

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

$$y^{r(\nu)} = \sum_{i=1}^{16} \sum_{\alpha=1}^{4} x_{i(\alpha)}^{r(\nu)} a_{i(\alpha)} \qquad \mathbf{1}$$

where r = 1, 2 and ν = 1, 2, ..., n_r. Coefficient of $a_{i(\alpha)}$ is a real number and is called 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_B^2 are given by

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

$$\sigma_B^2 = (1/N) \sum_{r=1}^{2} n_r (\bar{y}^r - \bar{y})^2 \qquad \mathbf{3}$$

where $N = n_1 + n_2$. To discriminate the sequences between the groups 1 and 2 most distinctly, we maximize the σ_B^2/σ^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). Sample score of any 16 nt sequence in the rat preproinsulin mRNA is then calculated by **1** together with the $a_{i(\alpha)}$ values. Our analysis demonstrates that the higher the score of a sequence, the stronger translation initiation signal the sequence has. In the next section, we analyze 16 nt sequences of the rat preproinsulin and its mutant mRNAs in terms of such sample scores.

MUTAGENESIS AROUND THE INITIATION CODON IN RAT PREPROINSULIN mRNA

Kozak used plasmid 255 as a shuttle vector which encodes and expresses the rat preproinsulin II gene (2,3). Initiation of transcription at the simian virus 40 (SV40) early promoter in the plasmid produces a chimeric mRNA with a 5′-untranslated sequence of ~128 nt: the first ~80 nt are encoded by SV40 DNA and the remainder comes from the rat preproinsulin gene (7). A *Hin*dIII site marks the boundary between SV40 and rat insulin DNA. For example, nucleotide sequence data of the chimeric mRNA for a B38 mutant are given in refs 2 and 3. A 16 nt sequence of mRNA around initiation codon (ATG) in the B38 mutant (AAAGCTTGGTTT/ATGT) lies at position 72/73, where the stroke (/) indicates the boundary between the non-coding region and the ATG codon, and where the number specifies the position of the sequence counted from the 5′-end of mRNA. In mutants of the B series, nucleotides flanking the ATG triplet were systematically varied. mRNA sequences of B31–B39 mutants are identical except for positions –3 and +4 around ATG, where the position of A is denoted by +1 (see Table 2). To determine how these base changes modulate translational efficiency, Kozak transfected the mutant plasmids into monkey (COS) cells, which were then incubated with [^{35}S]cysteine. Labeled proteins were extracted and analyzed by polyacrylamide gel electrophoresis. It was assumed that the observed variation in

proinsulin synthesis reflected the efficiency of translation initiation. Relative optical densities (OD) of the proinsulin product were measured in the mutants of B31–B39 series, and her experimental results were also summarized in Table 2 (3). The B38 mutant has T and T at –3 and +4 positions, respectively, and shows the weakest OD, while the B31 mutant with A and G at –3 and +4, respectively, shows the strongest OD. From the data in Table 2, the order of OD was found to be 5.0 of B31 mutant, 3.1 of B33, 2.6 of B35, 0.9 of B34 and B32, 0.7 of B39 and <0.2 of B38. We can see that only 2 nt changes at –3 and +4 positions varied the yield of proinsulin over a 20-fold range and that signal strength of the translation initiation should vary in accordance with the order of the magnitudes of OD.

Table 2. Comparison of experimental efficiency of translation initiation (relative OD) with calculated sample score of signal sequence in two series of rat preproinsulin and its mutant mRNAs

Mutant	Signal sequence	Relative OD[a]	Sample score
B 38	AAAGCTTGGTTTATGT	<0.2	1.8574
B 39	G T	0.7	3.6624
B 35	A T	2.6	3.9387
B 34	T G	0.9	2.2163
B 32	C G	0.9	2.7405
B 33	G G	3.1	3.9578
B 31	A G	5.0	4.2429
B137	AAGCTGCTTATTATGT	2.1/2.5	3.9118
B138	TTCTT	0.7/0.7	2.3464
B130	CCCCC	2.0/1.8	3.2588
B133	CCACC	4.1/5.2	4.7172
B140	TTTTT	<0.2	1.8220
B141	TATTT	<0.2	1.6955
B143	TATAT	<0.2	1.6393

[a]See ref. 3.

In another B series of B130–B143, pentanucleotides from –5 to –1 positions in the 5′-untranslated region of the mRNA were totally mutated. Table 2 also shows such sequence data together with the relative OD of proinsulin product (3). Among these mutants, B133 has a pentanucleotide of CCACC coincident to the consensus sequence in the region between positions –5 and –1, and is found to give the highest OD of 4.1–5.2. Mutants of B140, B141 and B143 possessing pentanucleotides of TTTTT, TATTT and TATAT, respectively, all of which differ from the consensus sequence, give the smallest OD (<0.2). On the other hand, B137 mutant has a pentanucleotide of TTATT, where only the –3 position of A agrees with the consensus, shows an intermediate degree of OD (2.1–2.5).

We attempted to explain those experimental results of translation efficiencies by the quantification analysis mentioned above. First, we studied the B31–B39 series of mutants. For example, the sequence data to be analyzed with B38 mRNA are demonstrated in Table 1. Quantification analysis gave the σ_B^2/σ^2 value as 0.8859; under this optimum condition, category weight values of $a_{i(\alpha)}$ were estimated as shown in Table 3. Here, the positive values of $a_{i(\alpha)}$ contribute greatly to the initiation signal, whereas the negative values are unfavorable for the signal. For example, $a_{i(\alpha)}$s with item i = 13 and category α = 1, with i = 14 and α = 4, and with i = 15 and α = 2 possess the largest values of 2.0069, 1.4145 and 0.9584, respectively, indicating that ATG at positions +1, +2 and +3 (they correspond to item i = 13–15, respectively) are essential for the initiation signal. The next important positions for the signal are i = 10 with α = 1 or 2, and i = 9 with α = 3, which show A or G at position –3, and C at position –4, respectively. These calculated results explain well features of the consensus sequence reported previously. As for the negative values of $a_{i(\alpha)}$, those of i = 13 with α = 2, 3, 4, of i = 14 with α = 1, 2, 3 and of i = 15 with α = 1, 3, 4 are the greatest. Such data imply that any nucleotide change within ATG destroys the signal in most cases (see Discussion). While the consensus sequence only gives the qualitative feature of the signal, our quantification analysis gives quantitative measures for positions and the kind of nucleotides which are favorable or unfavorable for the translation initiation signal.

Table 3. The optimum category weight values of $a_{i(\alpha)}$ for translation initiation signal calculated with quantification analysis of the data of Table 1[a]

Item (i)	Category (α) / nucleotide			
	1 / A	2 / G	3 / C	4 / T
1	–0.0644	–0.0506	0.1710	–0.1480
2	–0.0261	–0.0396	0.0325	0.0231
3	0.1840	–0.2429	0.1591	–0.1746
4	–0.1776	0.0963	0.0765	–0.0648
5	0.1035	0.1385	–0.0424	–0.1872
6	0.0321	0.0562	–0.0084	–0.0779
7	–0.1544	0.2263	–0.1415	–0.1195
8	0.1580	–0.1898	0.0890	–0.0581
9	–0.1583	–0.3946	0.3067	–0.1609
10	0.5079	0.2279	–1.0752	–1.6464
11	0.1291	–0.4332	0.0813	0.1071
12	–0.1064	–0.0842	0.1260	–0.1367
13	2.0069	–4.0432	–4.8030	–4.2920
14	–3.9129	–3.7112	–2.5774	1.4145
15	–1.9870	0.9584	–2.4603	–2.8596
16	–0.0131	0.2298	–0.2296	–0.2119

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

Sample score of any 16 nt sequence in B38 mRNA can be calculated with the category weight data of $a_{i(\alpha)}$ together with **1**, and score of the authentic translation initiation signal sequence (AAAGCTTGGTTT/ATGT) at position 72/73 is estimated to be 1.8574. In a similar way, quantification analyses were done with the remaining B31–B39 mutants, and the authentic signal sequences together with their calculated sample scores are summarized in Table 2. The order of sample scores for the 16 nt signal sequences is then compared with that of observed relative ODs. As was reported previously, the B31 mutant has the greatest

OD in the B31–B39 series, followed by B31 > B33 > B35 > B32, B34 > B39 > B38. In accordance with this, the authentic signal sequence of the B31 mutant has the largest sample score of 4.2429, followed by B31 > B33 > B35 > B39 > B32 > B34 > B38; the B38 mutant has the smallest score of 1.8574. The only discrepancy between orders of OD and sample score is that the score of B39 (3.6624) is larger than those of B32 (2.7405) and B34 (2.2163), while OD of B39 (0.7) is somewhat weaker than those of B32 (0.9) and B34 (0.9). These findings demonstrate that the greater the sample score of the signal sequence is, the more efficiently translation initiation occurs.

We examine the data of Table 2 in more detail. Kozak reported that comparison of relative OD of B35, B38 and B39 shows that, at position –3, A is more effective than G, and G is more effective than T (3). Comparison of B38 with B34, or B39 with B33, shows that G works better than T at position +4. These data agree well with our category weight data of $a_{i(\alpha)}$ in Table 3. At position –3 (i = 10), 0.5079 of A (α = 1) is greater than 0.2279 of G (α = 2), which is greater than –1.6464 of T (α = 4). Similarly at position +4 (i = 16), 0.2298 of G (α = 2) is greater than –0.2119 of T (α = 4). However, Kozak noted that the contributions of positions –3 and +4 are not simply additive. For example, G at position +4 enhanced ~5-fold with T at position –3 (B34 versus B39), 4-fold with G at position –3 (B33 versus B39) and only 2-fold with A at position –3 (B31 versus B35). Such non-additive contribution suggests interaction between the positions –3 and +4. Since our quantification analysis assumes independent contribution of item and category to sample score, no such interaction between two items (a context between two different positions within the signal sequence) is taken into consideration.

Next, quantification analyses were done with the B130–B143 series, and calculated sample scores of the authentic signal sequences were also compared with experimental efficiencies of translation initiation (relative OD) in Table 2. In both of the relative OD and sample scores, B133 mutant mRNA has the greatest values of 4.1/5.2 and 4.7172, while B143 has the least values of <0.2 and 1.6393, respectively. The order of relative OD values is found with B133 > B137 > B130 > B138 > B140, B141, B143, while B133 > B137 > B130 > B138 > B140 > B141 > B143 with sample scores. Again in the B130–B143 series, the order of relative OD values agrees well with that of sample scores.

The 3-fold increase in the relative OD between B138 and B130 and the 2-fold increase between B137 and B133 show that, instead of T, C at positions –5, –4, –2 and –1 enhances translation. These results correspond with the finding that all category weight values of C are larger than those of T at positions –5 (i = 8), –4 (i = 9) and –1 (i = 12) (see Table 3). At position –2 (i = 11), however, the value of C (0.0813) is somewhat smaller than that of T (0.1071), and T may be preferable from our quantification analysis. It was further noted from the relative OD data that C at position –3 apparently functions better than T (B138 versus B140). In mutants that lack A at position –3, the presence of A at position –2 or –4 does not compensate; i.e., while A at position –3 stimulates translation ~10-fold (B137 versus B140), mutants B141 and B143 translate only marginally better than B140. These experimental results are also understood in terms of our category weight values. At position –3 (i = 10), the value (–1.0752) of C is apparently greater than – 1.6464 of T, so that C functions better than T. At position –3 (i = 10), if A is substituted by T, a loss of sample score is as great as 2.1543. Although the presence of A at positions –2 (i = 11) and –4 (i = 9) gains scores of 0.0220 and 0.0026, respectively, they cannot compensate the loss of 2.1543. This gives a reason why A at position –3 stimulates translation greatly but why mutants B141 and B143 translate only marginally better than B140.

DISCUSSION

Comparison of experimental OD with calculated sample score in the series of B31–B39 and B130–B143 leads us to conclude that efficiency of translation initiation in vertebrate mRNA is predominantly determined by strength (sample score) of the 16 nt signal sequence. However, there may lie several questions relevant to our quantification analysis.

The first is how stable are the computed category weight values given in Table 3 and how they are applicable to initiation signals in other gene sets? This problem was discussed previously in the quantification analysis of splice signal sequences in mammalian mRNA precursors (5), where we examined if the category weight values might be unaltered, and recalculated the $a_{i(\alpha)}$ values using various genes to provide the members of group 2. The relative magnitudes of $a_{i(\alpha)}$ were found to be practically independent of the genes, so that our scoring system was strengthened. In the analysis of translation initiation signal, we used 699 signal sequences compiled by Kozak (1) to provide the members of group 1. Collection of such many signal sequences also contributes to the stability of the computed category weight values. The data in Table 3 are applicable to other genes, since the best translation initiation signal getting the largest sample score, as estimated by $a_{i(\alpha)}$s, agrees well with the canonical consensus sequence proposed by Kozak (1).

The second is whether 16 nt sequences that have the ATG at the right position are attributable to translation initiation sites and whether discrimination of initiation sites may be enhanced by excluding the invariant ATG. As is shown in the previous section, the B38, B140, B141 and B143 mutants exhibit very inefficient translation initiation and small sample scores of the 16 nt sequences which have the ATG at the right position. In such cases, initiation may occur not only at the first ATG codon but also at the downstream second ATG codon. Table 3 demonstrates that contributions of $a_{13(1)}$, $a_{14(4)}$ and $a_{15(2)}$ to sample score are so important that any nucleotide change within ATG may decrease the score greatly. Such a decrease can be compensated to some extent, if $a_{i(\alpha)}$ values at i = 1–12 and 16 gain positive scores. However, the total score of the 16 nt sequence possessing the non-ATG codon will not amount to sufficiently strong signal. This is confirmed by the finding that ribosomes can initiate translation at a non-ATG codon, such as ACG, CTG or GTG, in certain mRNAs but that initiation at non-ATG codons is usually inefficient and usually occurs in addition to using the first ATG condon (8). According to our quantification analysis, translation initiation reaction is described by its efficiency, which depends on the strength (sample score) of 16 nt signal sequence more or less different from the consensus sequence.

REFERENCES

1 Kozak, M. (1987) *Nucleic Acids Res.*, **15**, 8125–8148.
2 Kozak, M. (1984) *Nature*, **308**, 241–246.
3 Kozak, M. (1986) *Cell*, **44**, 283–292.
4 Kozak. M. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 2662–2666.
5 Iida, Y. (1987) *Comput. Appl. Biosci.*, **3**, 93–98.
6 Iida, Y. (1989) *Biochim. Biophys. Acta*, **1007**, 270–276.
7 Lomedico, P. T. and McAndrew, S. J. (1982) *Nature*, **299**, 221–226.
8 Kozak, M. (1991) *J. Cell Biol.*, **115**, 887–903.