Data Mining of Structural Features in 5' Untranslated Regions of Cellular mRNAs

Shu-Yun Le and Jacob V. Maizel, Jr,
Laboratory of Experimental and Computational Biology
NCI Center for Cancer Research, National Cancer Institute, NIH.
Bldg. 469, Room 145, Frederick, Maryland 21702, USA

Abstract *Statistical analyses of current databases show that approximately 16% of the vertebrate mRNA 5' untranslated regions (5'UTR) are over 300 nucleotides (nt) long. The important role in translational control for long 5'UTRs containing multiple silent upstream AUG (uAUG) triplets is now widely accepted. Experimental studies have revealed that a special sequence, termed the internal ribosome entry site (IRES) allows the translational machinery to skip over the uAUG, interact with it and start translation at the true initiator. Interestingly, a common Y-shaped stem-loop followed by a short complementary sequence to the 3' end of human 18S rRNA immediately upstream to the initiator has been found in cellular IRESs by an integrated method composed of a set of computer programs. These Y-shaped structures were used to search cellular mRNAs from tumor-associated proto-oncogenes and cell factors with the designed pattern. We discovered a common structure motif in various 5'UTRs of cellular mRNAs encoding oncoproteins and cell factors related to cell proliferation. These 5'UTRs are long and have high G+C content and/or multiple uAUGs. This structural motif is suggested to play a functional role in the IRES-mediated initiation of cellular mRNAs. Our computational tools are suitable for finding common structures in nucleic acid sequences.*

Keywords: Eukaryotic 5'UTR, Y-shaped stem-loop, 18S rRNA-complementary sequence, IRES.

1 Introduction

Messenger RNA (mRNA) has long been recognized as the immediate source of information for translation from sequences in the four base alphabet of nucleic acids to the twenty amino acids of proteins. A mRNA may also have 5'-untranslated region (UTR) before and 3'UTR following the coding region. In most mRNAs, translation into protein begins at the first initiation (AUG) codon by ribosomal 40S subunits that are recruited at the capped 5' end of mRNAs. Translational initiation by the cap-dependent ribosomal scanning is highly efficient in mRNAs that have short, unstructured 5'UTRs. A notable exception of this phenomenon exists in picornaviruses and certain cellular mRNAs [1]. Among them multiple silent upstream AUGs (uAUGs) exist in the 5'UTR, yet translation starts from an AUG more than hundreds of nucleotides (nt) downstream. Experimental study [1-4] indicates that there are cis-acting internal ribosome entry sites (IRESs) in their 5'UTRs by which 40S subunits can alternatively enter into the mRNAs and initiate translation in the initiator by skipping over the multiple uAUGs.

Cellular mRNAs with reported IRESs cover a wide range of eukaryotic sequences including oncoproteins, growth factors, and immune or inflammation mediators that often have a long, G+C rich and/or structured 5'UTR with multiple uAUG [1]. The highly structured 5'UTR is hypothesized to provide a built-in blockade against efficient classical cap-dependent ribosomal scanning. Although it is not clear why some cellular mRNAs employ internal initiation rather than the classic, cap-dependent initiation, possible roles for this kind of translational regulation are alternative expression in developmental differentiation and response to stress and environmental changes [1]. Internal ribosome binding depends on interactions between a *cis*-acting IRES and *trans*-acting protein factors. The RNA-protein interaction involves the specific recognition of sequences and structural motifs of the IRES elements by the proteins. There is no notable sequence and structure homology among cellular IRES elements or between viral and cellular IRES elements. Although computational analyses revealed that some cellular IRES elements can form a Y-shaped 3-way junction of stem-loops close to the initiator AUG [5], structural features of most cellular IRES elements are largely unknown. Do the functional IRESs in 5'UTRs comprise an undiscovered, structural motif that could be crucial in the translation control? Will an extensive search for the common structural motif find them in other cellular mRNAs with similar biology? Identification of conserved structural features in the 5'UTR of cellular mRNAs is essential and will advance a better comprehension of the properties of cellular IRESs.

In this study we first determine the common structural motif in the reported cellular IRES elements by an integrated method using a set of computer programs. Computation of RNA folding and extensive statistical analyses are used to discover local well-ordered RNA secondary structures in the 5'UTRs. Common Y-shaped structural motifs are inferred by structure analysis and comparison. We show that the Y-shaped 3-way junction of stem-loops followed by a short complementary sequence to the 3' end of human 18S rRNA immediately upstream from the AUG triplet is a conserved structural feature for the cellular IRES elements. The characterized Y-shaped stem-loop is used to design a structural pattern to search for more cellular mRNAs from tumor-associated proto-oncogenes and signal transduction genes by the program RNAMOT [6] and the integrated method. A typical property of these mRNAs is their long 5'UTRs that contain multiple uAUG triplets or high G+C content. A functional role of the detected structure motif in the regulation of translational control is suggested. We expect that finding additional sites will further our understanding of the cap-independent translational mechanism, and reveal important features of structure and function in the 5'UTR of cellular mRNAs.

2 Methods

Our goal is to use computational tools to discover potential functional elements and structural features in the 5'UTRs and to correlate them with known experimental properties. The theoretical basis for our method (Fig. 1) lies in the hypothesis that the functional structured RNA (FSR) elements possess well-ordered conformations that are both thermodynamically stable and uniquely folded. [7-9]. As suggested previously [10], FSR elements are often closely associated with the local, unusual folding regions (UFRs) in the sequence. The method includes a set of computer programs,

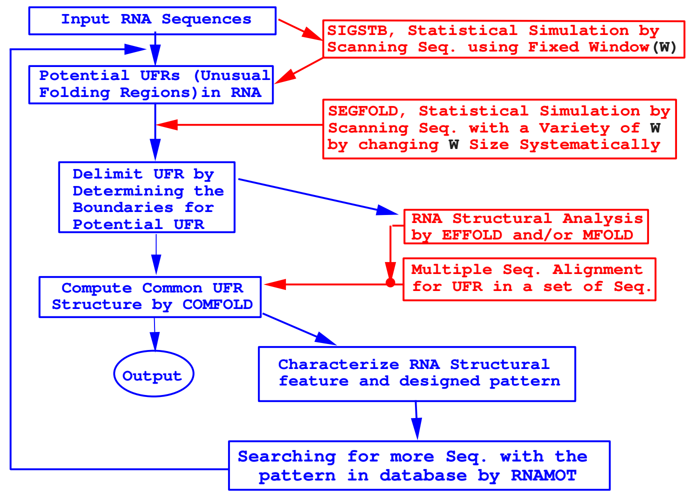

Figure 1: Flowchart of data mining of structure feature in RNAs. For details of the programs SIGSTB, SEGFOLD [10], EFFOLD [11], MFOLD [14], COMFOLD [12] and RNAMOT [6] see previous publications.

SEGFOLD (or its modified version SIGSTB) [10], EFFOLD [11], COMFOLD [12] and RNAMOT [6].

In the first step, we search for UFRs in the cellular 5'UTRs that contain IRES elements determined by extensive experimental studies [1]. The program SEGFOLD or SIGSTB is used to explore a sequence by choosing successive segments of the mRNA and comparing the lowest free energy of the actual sequence to a number of randomly shuffled sequences of the same size and base composition. Simultaneously, the comparison is also made between the lowest free energy of each segment and the average of all segments resulted from the scanning. The lowest free energy of the folded segment is calculated by a dynamic programming algorithm [13-14] and Turner energy rules [14]. The two z-scores, SIGSCR and STBSCR are used to quantitatively measure the statistical significance and property of the local thermodynamic stability, respectively. When a local UFR is found by a fixed window, the boundary of the UFR is delimited by a more extensive search in which the window is systematically changed with a range of sizes using SEGFOLD.

We construct a common high-ordered structure for the detected UFRs based on the analysis of RNA folding and structure similarity [11-12]. In the

Eukaryotic mRNAs	Total Number	Number of Sequences that 5'UTR size >= 100	200	300	600	Mean of 5'UTR Length (nt)
Human	6669	3725	2055	1246	415	207
Other mammal	2106	833	348	202	56	138
Rodent	7056	3844	2048	1165	345	192
Other vertebrate	2846	1265	610	311	81	152
Invertebrate	4054	1975	1107	709	268	205
Plant	5993	1826	592	298	76	106
Fungalt	931	406	215	153	63	174

Table 1: Distribution of the 5'UTR Length for mRNA Sequences Collected in the UTR-DB [15].

step, we generally need some manual manipulations to infer a common structure. For the last step we characterize the basic structural feature from the computed common RNA structures and design the search pattern. The program RNAMOT [6] is used to search for more 5'UTR sequences to see if the basic pattern of the structure motif can be detected in the sequence database. For detected sequences we repeat the previous steps and determine the optimized well-ordered RNA structure based on both the structure similarity and thermodynamic stability.

3 Results and Discussion

All 5'UTRs of eukaryotic mRNAs are from a specialized database UTRdb [15]. Statistics of the 5'UTR lengths are listed in Table 1 and indicate that almost 16% of vertebrate 5'UTRs are 300 or more nucleotides (nt). Multiple uAUG are found in the long 5'UTR of vertebrate mRNAs and the distribution is shown in Fig. 2.

Statistical analyses indicate that approximately 26% of human mRNAs have 2 or more uAUGs and 10% human mRNAs have 5 or more uAUGs in the 5'UTR. Also, 5% of other mammalian mRNAs have 5 or more uAUGs in their 5'UTR. It has been shown that a large number of 5'UTRs of oncoproteins, growth factors, transcription factors, signal transduction genes, receptors and human disease genes have this distinct feature and high G+C content [16]. We expect that there are a growing number of interesting cases where internal initiation may play an important role in the translational regulation of these cellular mRNAs.

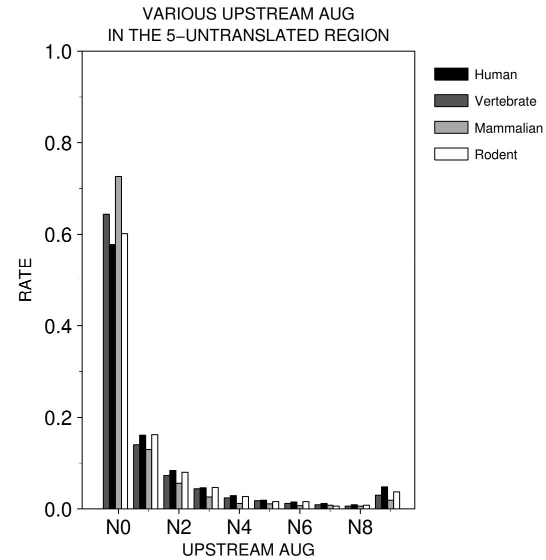

Figure 2: Distribution of upstream AUG (uAUG) triplet in the 5'UTR of human, other mammalian, rodent and other vertebrate mRNAs. In the figure N0 represents that there is no uAUG in the 5'UTR. N1-N8 represent 1-8 uAUG found in the 5'UTR, respectively. N9 means 9 or more uAUGs found in the 5'UTR.

3.1 Common RNA structural motif of IRES elements

The proposed RNA structural motif found in the 5'UTRs that contain cellular IRES elements are listed in Table 2. We first searched for UFRs in the 5'UTRs by the program SEGFOLD. Human eIF4G mRNA contains 4 uAUGs and 58% G+C content in the 368-nt 5'UTR. Using a fixed 100-nt window we detected a UFR, fragment 219-318, in the 5' portion (1-1000) of the mRNA. The computed SIGSCR and STBSCR of the UFR are -4.10 and -2.22, respectively (Fig. 3). It means that the thermodynamic stability of the distinct fragment is about 4.10 standard deviations (std) more stable than by chance, and 2.22 std more stable than other 901 overlapping fragments of 100-nt in the tested eIF4G sequence. Lower SIGSCR highlights the high statistical significance of the distinct UFR. The computed RNA secondary structure of the UFR is a Y-shaped three-way junction of stem-loops A, B, and C (Fig. 4). There is a 8-nt complementary sequence to the 3' end of human 18S rRNA (1823-1869) between the Y-shaped structure and the translational initiator AUG. Those comple-

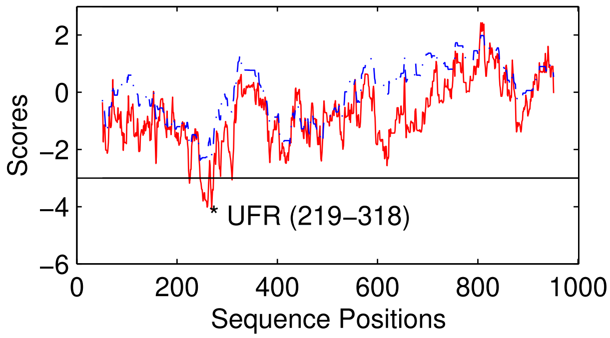

Figure 3: Significance score (SIGSCR) and stability score (STBSCR) against the midpoint of the window of 100-nt along the mRNA (1-1000) of human eIF4G mRNA. The SIGSCR and STBSCR are displayed by red solid line and blue dash line, respectively. The detected UFR in the 5'UTR is located at the region 219-318.

mentary sequences observed in human 18S rRNA are located at the upstream and downstream single-stranded regions (1836-1838 and 1861-1865) of the folded hairpin structure (1839-1860) in the 3'-end of the rRNA. The analogous structural motifs are also found in the 5'UTR of BiP, FGF-2, PDGF, and AML1 mRNAs (Table 2) that all include a IRES element in the 3' portion of the 5'UTR [1].

A similar structural feature was also found in the 5'UTR of human vascular epithelial growth factor (VEGF). Human VEGF contains a very long (1038 nt) 5'UTR that is marked by a high G+C content (83%) and one uAUG. The UFR (845-987) was detected by the 143-nt window and computed SIGSCR and STBSCR are -2.88 and -2.08, respectively. The common RNA structure is also a Y-shaped stem-loop structure. However, there is an additional small stem-loop D downstream of the Y-shaped structure, which is further followed by a 6-nt 18S rRNA-complementary sequence close to the translational initiator. The location of the complementary sequence in human 18S rRNA is almost same as that we observed in the case of human eIF4G (see Figs. 4-5). Mutational analyses demonstrate that the 294 nt segment (745-1038) in the 3'-end of human VEGF 5'UTR contains a very potent IRES activity [17]. The structural feature is con-

5'UTR (human)	Size (nt)	No. of uAUG	Y-shaped structural motif A	B	C	D	Complementary Sequence
AML1	1580	15	1469-1480 //1546-1558	1482-1520	1521-1545	N	CUUGUUGUG-(0 nt)AUG
BiP	221	0	129-142 //183-193	144-160	161-179	N	ACuGGCU-(6 nt)AUG
c-myc	513	0	228-249 //340-368	252-311	313-336	Y	UGcUUAGAC (1 nt)CUG
eIF4G	368	4	219-236 //301-317	237-259	260-296	N	GAUCCaaACC-(29 nt)AUG
FGF-2	466	3	204-221 //263-276	222-241	242-259	N	GCGGCU-(5 nt)CUG
			355-369 //420-437	371-391	392-418	N	GGgGAUCCcg-GCC(16 nt)AUG
PDGF2/ c-sis	1022	3	941-944 //992-995	946-969	970-990	N	GCCcggaguCGGC-(0 nt)AUG
VEGF	1038	1	845-855 //978-987	858-907	909-977	Y	GGCCUCC-(6 nt)AUG

Table 2: Y-shaped structural motif and a short complementary sequence to the 3' end of human 18S rRNA sequence found in the cellular 5'UTR that contains cellular IRES. The folding regions of stems A, B, and C in the Y-shaped motif are listed in the columns 4, 5 and 6. An additional stem-loop D between the Y-shaped motif and 18S rRNA-complementary sequence is denoted by letters Y (Yes) and N (No) in the seventh column. The 18S rRNA-complementary sequences are represented by capital letters in the last column. All of complementary sequences observed in human 18S rRNA are located at the upstream and/or downstream single-stranded regions (1823-1838 and 1861-1869) of the folded hairpin structure (1839-1860) in the 3'-end as shown in Figs. 4 and 5. The codon CUG is an alternative initiator in FGF-2 and c-myc mRNAs

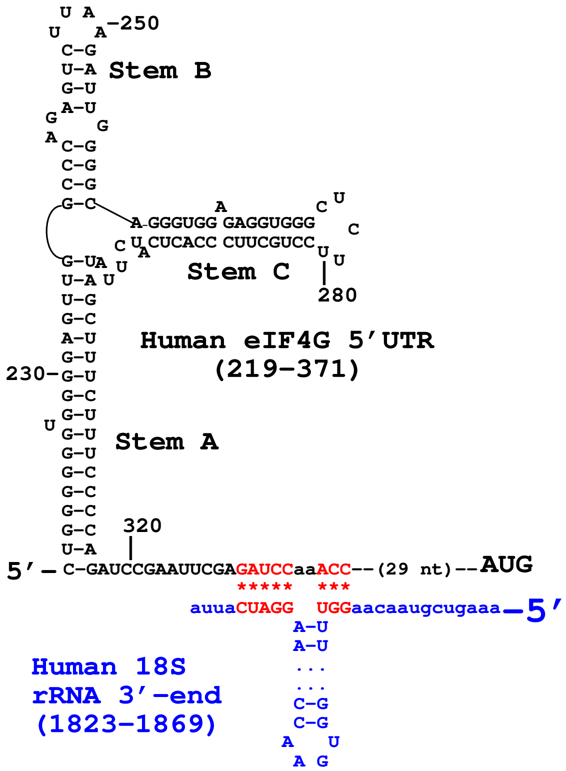

Figure 4: Y-shaped structure computed in the 3' portion of the 5'UTR of human eIF4G mRNA.

served in the 5'UTR of bovine and mouse VEGF. An analogous structure is also determined in the 3' end of the 5'UTR of c-myc oncoprotein mRNA. We suggest that the Y-shaped stem-loop (A-C), with or without a small stem-loop (D), followed by a short complementary sequence to the 3' end of 18S rRNA close to the translational initiator together is a common structural motif of cellular IRES elements.

3.2 Data mining of cellular 5'UTR

To search for similar structural feature in the UTRdb with a designed pattern we need to characterize the basic feature from the derived structural motifs of cellular IRES. The basic feature in the designed pattern included the minimal and maximal base-pair size of stems A-C, the mismatching allowed in each stem, the ranges of loop sizes and distance between the Y-shaped structure and the codon AUG. In practice, the designed pattern was defined quite flexibly and the program RNAMOT was functioned as a filter in the data mining. Our primary target was focused on those cellular mRNAs encoding oncoproteins and cell factors related to cell proliferation. Once a candidate of 5'UTR was found, extensive data mining was employed to discover the possible common structure by the computer program SEGFOLD, EFFOLD and COMFOLD based on both the thermodynamic stability and morphologic similarity of RNA structure. The preliminary result was summarized and listed in Table 3. The common structure found in the 3' portion of the 5'UTR of the various cellular mRNAs is similar to the IRES elements found in eIF4G and VEGF. Also, complementary sequences observed in human 18S rRNA are all located at the upstream and/or downstream single-stranded regions (1823-1838 and 1861-1869) of the hairpin structure (1839-1860) in the 3'-end. These mRNAs contain multiple uAUGs in the long 5'UTR sequence that often hinder the classical scanning of translational machinery so that the translation is inefficient in the normal condition. Our results raise the possibility that internal initiation may be used in the regulation of the translational control of these cellular mRNAs to improve the gene expression when they are needed in cell development. The distinct feature represented in the common structural motif would have a general function involved in the RNA-protein interaction of the internal ribosomal binding mechanism of the translational control. Our data are useful to help experimental design for discovering more cellular IRES elements in mRNAs encoding the important proteins that are produced as a response to a variety of stress situations, such as angiogenesis and inflammation, during apoptosis, and to be correlated with the pathogenesis of human diseases.

4 Conclusion

We have discovered a common structural motif in the 3'-end of the 5'UTRs that contain a cellular IRES element by computational methods. The common structural motif is delineated to be a Y-shaped stem-loop with or without an attached stem-loop, which is further followed by a short complementary sequence to the 3' end of 18S rRNA (1823-1869). The conserved structural feature characterized from cellular IRES allowed us to search for the distinct structural profile in the sequence database by means of pattern recognition, analyses of RNA folding and thermodynamic stability, and statistical simulation. Our preliminary results show that various cellular mRNAs contain the structural feature in the 3'-end of the 5'UTR. A common structural motif is suggested to be associated with the important role of reported IRES. We will continue our efforts to refine the common structural motif and to find more cellular mRNAs comprising the distinct structure in the 5'UTR.

References

[1] Hellen, C.U.T. and Sarnow, P., Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes & Development* **15** (2001) 1593-1612.

[2] Pelletier, J. and Sonenberg, N., Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334** (1988) 320-325.

[3] Macejak, D.G. and Sarnow, P., Internal initiation of translation mediated by the 5' leader of a cellular mRNA. *Nature* **353** (1991) 30-94.

[4] Vagner, S., Gensac, M.C., Maret, A., Bayard, F., Amalric, F., Prats, H, and Prats, A.C., Alternative translation of human fibroblast growth factor 2 mRNA occurs by internal entry of ribosomes. *Mol. Cell. Bio.* **15** (1995) 35-44.

[5] Le, S.-Y and Maizel Jr., J.V., A common RNA structural motif involved in the internal initiation of translation of cellular mRNAs. *Nucl. Acids Res.* **25** (1997) 362-369.

[6] Laferriere, A., Gautheret, D. and Cedergren, R., An RNA pattern matching program with enhanced performance and portability. *Computer Appl. Biosci.* **10** (1994) 211-212.

[7] Draper, D.E., Strategies for RNA folding, *Trends Biochem Sci* **21** (1996) 165-169.

[8] Schultes, E.A., Hraber, P.T. and LaBean, T.H., Estimating the contributions of selection and self-organization in RNA secondary structure. *J Mol. Evol.* **49** (1999) 76-83.

[9] Gutell, R. R., Comparative studies of RNA: inferring higher-order structure from patterns of sequence variation. *Curr. Opin. Struct. Biol.* **3** (1993) 313-322.

[10] Le, S.-Y., Chen, J.-H. and Maizel Jr., J.V., In Sarma, R.H. and Sarma, M.H. eds. *Structure & Methods: Human Genome Initiative and DNA Recombination.* Vol. I, pp 127-136, (Adenine Press, Schenectady, 1990).

[11] Le, S.-Y., Chen, J.-H. and Maizel Jr., J.V., Prediction of alternative RNA secondary structures based on fluctuating thermodynamic parameters. *Nucl. Acids Res.* **21** (1993) 2173-2178.

[12] Le, S.-Y, Zhang, K., Maizel, Jr. J.V., A method for predicting common structures of homologous RNAs. *Comp. Biomed. Res.* **28** (1995) 53-66.

[13] Zuker, M., On finding all suboptimal foldings of an RNA molecule. *Science* **244** (1989) 48-52.

5'UTR (human)	Size (nt)	No. of uAUG	Region of Y Motif	Stem D	Complementary Sequence
abl M14753	340	6	225-321	N	GGU--ACC-UAUUAUuACUUU -(0 nt)AUG
c-abl M14752	364	0	276-361	N	-
bcr X02596	488	2	373-455	N	GGCGG--CGC(9 nt)CGGC -(6 nt)AUG
c-erb Y00479	333	3	117-249	Y	GGcAUCC(9 nt)UUGaa- GUGA-(0 nt)AUG
c-erbA-1 X55005	466	4	258-382	Y	GGcAUCC(9 nt)UUGaa- GUGA-(0 nt)AUG
IL-15 X91233	316	10	221-290	N	UAAgGAUUUACC--GU ----GGCUUU-(5 nt)AUG
Int-2 X14445	491	3	396-456	Y	GAUGCC-(3 nt)AUG
mas M13150	267	3	117-249	N	CCaACCU-GaGGCcU- (4 nt)AUG
mos J00119	479	1	369-473	N	AUcAUC-(0 nt)AUG
(mouse or others)					
abl-2 U13835	144	3	55-125	N	UCCaGCCUcCGAC -(0 nt)AUG
abl-3 X07539	219	0	90-166	N	UAA-GGUCCuugugaGCC- acgUUGUGGU-(25 nt)AUG
abl-4 X07541	1168	11	1059-1145	N	CACCUaUUAUuGCUUU- (0 nt)AUG
BiP M14866 (rat)	206	0	114-169	N	CCGCUgagcgACuGACU- (19 nt)AUG
BiP M17169 (hamster)	150	0	64-125	N	GGCCCACagcGCcGGC- (3 nt)AUG
Int-2 Y00848	864	3	776-833	Y	GAUGCC-(3 nt)AUG
FGF-2 M22427 (rat)	532	0	310-375	N	GUCCgGCU-(8 nt)CUG
	(CUG is an alternative initiator for FGF-2)				
			438-510	N	GUCCcgggGCC- GCGG-(7 nt)AUG
c-myc Y00396 (rat)	413	0	71-222	Y	UUAUU-UGA(3 nt)CUG
	(CUG is an alternative initiator for c-myc)				
mas U96273	341	5	86-212	N	CACCg-(0 nt)AUG
mos X12449 (green monkey)	479	1	369-473	N	AUcAUC-(0 nt)AUG
mos X52952 (rat)	482	2	369-477	N	UAAUc-(0 nt)AUG
mos M19412 (chicken)	487	4	377-481	N	AUcAUC-(0 nt)AUG
mos X13311 (Xenopus)	483	2	373-477	N	AUcAUC-(0 nt)AUG
VEGF U41383	1014	1	818-962	Y	AcGGcCU-CC(6 nt)AUG
VEGF M32976 (bovine)	533	0	351-494	Y	A-GGcCU-CC(6 nt)AUG
eIF4F L16923 (Yeast TIF4631)	528	11	438-507	N	ACCUaUUAC(4 nt)AUG
eIF4F L16924 (Yeast TIF4632)	528	4	437-502	N	AAUaGAUCaaUUGU-Ag- GcACU-(0 nt)AUG

Table 3: Y-shaped structural motif and a short 18S rRNA-complementary sequence found in the other cellular 5'UTRs. The folding regions of stems A, B, and C in Y-shaped motif are listed in the columns 4, 5 and 6. The complementary sequences to 3'-end of human 18S rRNA are represented by capital letters in the last column. All of complementary sequences in human 18S rRNA are located at the upstream and/or downstream regions (1823-1838 and 1861-1869) of the hairpin structure (1839-1860) in the 3'-end as shown in Figs. 4 and 5. The accession number of the sequence in Genbank is listed in the first column.

[14] Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H., Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.* **288** (1999) 911-940.

[15] Pesole, G,, Liuni, S,, Grillo, G,, Ippedico, M,, Larizza, A,, Makalowski, W,, and Saccone, C., UTRdb: a specialized database of 5' and 3' untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **27** (1999) 188-191.

[16] Kozak, M., An analysis of vertebrate mRNA sequences: intimations of translational control. *J. Cell Biol.* **115** (1991) 887-903.

[17] Akiri, G., Elroy-Stein, O., Nahari, D., Finkelstein, Y., Le, S.-Y. and Levi, B.Z., The 5' Untranslated region (5'UTR) of vascular endothelial growth factor (VEGF) contains an internal ribosome entry site (IRES) and promoter activity. *Oncogene* **17** (1998) 227-236.

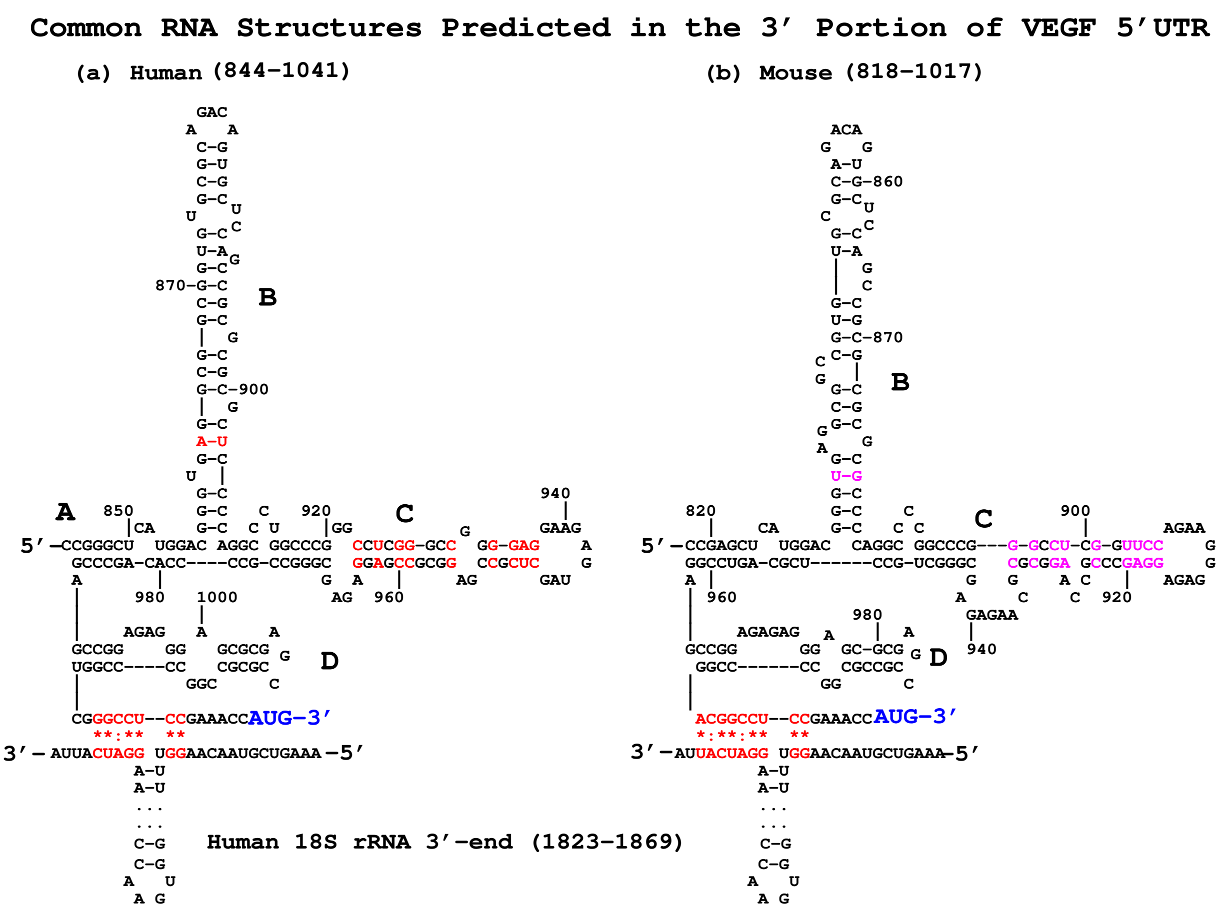

Figure 5: Y-shaped structure computed in the 3' portion of the 5'UTR of VEGF mRNA.