Finding Double-stranded RNA Structures in the 3' Untranslated Region of Eukaryotic mRNAs

Shu-Yun Le and Jacob V. Maizel, Jr
Laboratory of Experimental and Computational Biology, CCR
National Cancer Institute, NIH, Bldg. 469, Room 151, Frederick, MD 21702
Tel: 301-846-5576; Fax: 301-846-5598, e-mail: shuyun@ncifcrf.gov

Abstract

Recent development in the study of RNA silencing indicates that double-stranded RNA (dsRNA) can be used in eukaryotes to block expression of a corresponding cellular gene. In the RNA interference (RNAi) pathway, dsRNAs serve as the initial trigger that are chopped by an ribonuclease termed "Dicer" and result in mRNA degradation and aberrant. It has reported that a large dsRNA structure in the 3' untranslated region (3'UTR) may correlate with the translation suppression. In this study, we search for dsRNA structures in the UTR database. The occurrence rate of the large dsRNA structure in 3'UTRs ranges from 0.01% in plant to 0.30% in vertebrate mRNAs. However, small dsRNAs of 21-23 bp are much more than the large one. These dsRNAs are very significant in Monte Carlo simulations and are uniquely well-ordered. The detected dsRNA structure in the 3'UTR of *Drosophila bicoid* is in good agreement with the *cis*-acting element that plays a key role in the translational repression and localization. Our data mining of dsRNAs can be used to explore possible RNAi in the RNA-based regulation of gene expression.

Introduction

It is conceivable that the 3' untranslated region (3'UTR) is not traversed by ribosomes [1]. Therefore 3'UTR seems a place for the assembly of complexes that can contribute to the overall regulation of cell growth, mRNA metabolism and localization by post-transcriptional regulation [1-3]. The response element in 3'UTRs can control the function of eukaryotic mRNAs by affecting the polyadenylation, stability, localization and translation of the mRNAs. This type of regulations of the gene expression are particularly common in the early development, when transcription is inactive, or nearly so, and the post-transcriptional regulation is needed to effect rapid changes in protein levels to regulate key developmental decisions. Although the regulatory mechanism of the 3'UTR is still not well understood, increasing evidence from several laboratories indicates that specific sequences, RNA structural motifs and double-stranded RNA (dsRNA) within the 3'UTR are actively involved [1-5, 8] in the post-transcriptional mediation.

Cytoplasmic localization of mRNAs is targeted to the 3'UTR by *cis*-acting RNA elements and *trans*-acting factors [5]. The regulation signal of localized translation in the 3'UTR usually contains a large RNA segment predicted to be rich in the dsRNA stem-loops [6]. Recent advances in studies of RNA interference (RNAi) indicate that RNA silencing can be experimentally activated by dsRNA through its sequence-specific interaction with the target segment in the 3'UTR [7, 8]. It is clear that the 3'UTR is the place for harboring a wide range of response elements important in the regulation of gene expression rather than a junk sequence without any biological functions as thought many years ago. Accumulated data indicate that the dsRNA structure in the 3'UTR is involved in the translation repression [6, 9-11]. In eukaryotic cell, the long dsRNA can be chopped to form smaller 21-23 base-pairs (bp) short interfering RNAs (siRNA) containing about 19 bp duplexes by an RNaseIII-like enzyme named Dicer [12]. These siRNAs can guide a multi-component nuclease complex, which specifically identify and degrade target mRNAs and lead to the post-transcriptional gene silencing. In light of the recent evidence, an intriguing question is if the natural, endogenous dsRNA in 3'UTRs could perform a similar inhibition of gene expression under some conditions. In this study, we search for the dsRNA structure in the UTR database. The detected dsRNA structures are both statistically significant and well-ordered. We also found one of the detected dsRNA structures is in good agreement with the *cis*-acting element that play a key role in the localization and translation of *Drosophila bicoid*

mRNA [6]. We suggest that the dsRNA structure in the 3'UTR with highly statistical significance may correlate with the RNA-based regulation of gene expression.

Methods

All 3'UTR sequences (3'UTRs) were from a specialized database of UTRdb (release 7.0, May 1998) [13]. The 3'UTRs used in this study were divided into seven eukaryotic groups, namely (1) 7503 human 3'UTRs; (2) 2457 other mammal 3'UTRs; (3) 7633 rodent 3'UTRs; (4) 3499 other vertebrate 3'UTRs; (5) 5067 invertebrates; (6) 8116 plants; and (7) 1154 fungi. In the data mining of dsRNA structure we used an integrated approach of statistical and computational tools of RNA folding and pattern search. Among them, the program SEGFOLD [14] and EDscan [15] were used to evaluate those unusual folding and well-ordered structures. The program RNAMOT [16] was used to search for similar structural patters in the UTRdb with the designed pattern. The structural pattern was designed based on the large stalk-like dsRNA found in onconase 3'UTR [11] in which the minimal size of the base-pairing in the dsRNA was allowed to be 50 bp with less 4 mismatches. Based on the preliminary data from pattern searches, structural analysis was further continued by SEGFOLD [14] and EFFold [18] to determine the robust structure models and the statistical significance for these long dsRNA structures.

The small dsRNAs were searched by StemED, a new version of EDscan [15] that was specifically designed to identify distinct stem-loops efficiently. The predicted stem-loops were then evaluated and statistically inferred by SigED [21]. The detected small dsRNAs contained 21 bp at least and with an allowable one mismatch in the stem.

Results and Discussion

Large dsRNA Structures in 3'UTR Sequences

The distinct large dsRNA structures were detected in the database by using a cutoff of $SIGSCR < -6.5$ and summarized in Table 1. We found only 14 long dsRNA structures from 7503 sequences of human 3'UTR. For 14 mRNAs that had large dsRNA, their 3'UTR lengths ranged from 655 nt to 4148 nt. The largest duplex included 107 consecutive bp and was found in the 3'UTR of human aurora/IPL1-related kinase (AIK). The largest dsRNA structure was detected in human mRNA for KIAA0186 gene and the total size of dsRNA stem was 278 bp. We also found 3 sequences that encompassed the dsRNA structure in 2457 mammalian mRNAs, 4 sequences in 7633 rodent mRNAs, 11 sequences in 3499 other vertebrate 3'UTR, 7 sequences in 5067 invertebrate 3'UTR, and only 1 sequence in either 8116 plant or 1154 fungi 3'UTRs. The distinct large dsRNA structures were found at very low frequencies and were ranged from 0.01% in plant up to 0.30% in vertebrate 3'UTR. Among the detected long dsRNA structures, the 3'UTR of *Florida lancelet (F.lan)* mRNA (1552 nt) had the largest stalk-like dsRNA containing 402 bp in the stem, and the largest duplex included 227 consecutive bp.

In our search, we also found some predicted dsRNA structures that were composed of repeated dinucleotide sequences, such as GU, GC and AU. However, the simple RNA dsRNAs were not uniquely folded [15], and had low statistical significance in the Monte Carlo simulation and were filtered out in our data mining. Those small dsRNAs found in the database are not listed in Table 1, and can be accessed by a request.

Statistical Significance of dsRNA Structures

The computed SIGSCR and STBSCR [14] of these dsRNA structures are also listed in Table 1. These detected dsRNA structures had very large SIGSCR in negative numbers. For example, the SIGSCR of the dsRNA 1477-1804 (99-426 in the 3'UTR) in human AIK mRNA was -39.84 (Fig.1). This is statistically very significant. It means that the dsRNA structure is about 39.8 standard deviation (std) more stable than by chance and implies that the dsRNA structure can not be expected to occur by chance. The average and std of the SIGSCR values of those segments having the same size as the segment 1477-1804 were -4.03 and 9.69 in the AIK mRNA. It indicated that the dsRNA structure was also more statistically significant than other segments in the complete AIK mRNA. The expected probability of such unusual dsRNA in the AIK mRNA was less than 0.0001 in the normal approximation to the statistical test. The STBSCR of the dsRNA structure was -3.81 and implied that the dsRNA stem-loop was highly stable than other segments in the simulation of RNA folding, although the rate of G and C in the unusual folding region (UFR) was about 41%. For these mRNA sequences listed in Table 1, the mean and std of STBSCRs computed in each sample were 0 and 1, respectively.

RNA structural analyses of the 3'UTR sequences by EFFold indicated that there were no other alternative structures that was significantly different from the predicted dsRNA stem-loops in searching for possible alternative predictions. The predicted

dsRNA stem-loops were 'well-determined'. The feature of the well-determined structure can also be indicated by 'energy dot plots' of Mfold [18].

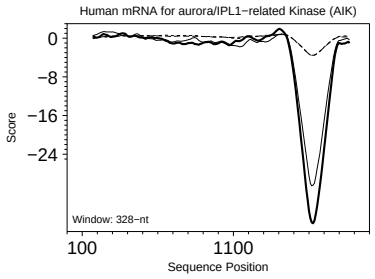

Figure 1: Distributions of the significance score (SIGSCR, continuous curve) and stability score (STBSCR, broken curve) in human aurora/IPL1-related kinase (AIK). In the plot, two scores were computed by both the Turner [18] and Tinoco [22] energy rules and were represented by thick and thin curves, respectively. All score data were averaged over 9 successive segments and then smoothed by five successively overlapping segments. The detected UFR is located at 1477-1804 in human AIK mRNA (99-426 in the 3'UTR). The profile was obtained by plotting the SIGSCR and STBSCR of each segment window against the position of the middle nt in the segment by sliding the window throughout the sequence. The two scores are defined as follows: $SIGSCR = (E - E_r)/std_r$ and $STBSCR = (E - E_w)/std_w$, where E is the lowest free energy of formation for RNA folding of a specific RNA fragment, E_r and std_r are the mean and standard deviation (std) of the lowest free energies computed from 500 random shuffling RNA fragments; and E_w and std_w are the mean and std of the lowest free energies resulting from sliding a window of 328-nt along the sequence from 5' to 3'-end. To speed computation, E_r and std_r were computed by a set of coefficients that were derived from the least-squares fit to the 500 random shuffling sequences.

Structural Analysis in the 3'UTR of *Drosophila bicoid* mRNA

Monte Carlo simulations also detected a 551-nt dsRNA stem-loop (1814-2364) in the *Drosophila bicoid* (*bcd*) mRNA (Accession No. X14458). The distinct UFR was 175-nt downstream of the stop codon (1637-1639). The SIGSCR of the 551-nt UFR in the *bcd* 3'UTR (175-725) was -10.28 (see Fig. 2). The transgenic analysis mapped a 625 nt region in the *bcd* 3'UTR that was required for all steps of localization process [13]. The core region of the *cis*-acting element, segment 181-720 (1820-2359 in mRNA) of the *bcd* 3'UTR, was a response element of *staufen* protein that was a *trans*-acting factor required in the localization of the *bcd* mRNA [19]. It suggests that the detected large dsRNAs may encompass such RNA functional elements in which the folded RNA duplex plays a crucial role in the translational control, rather than its primary sequence.

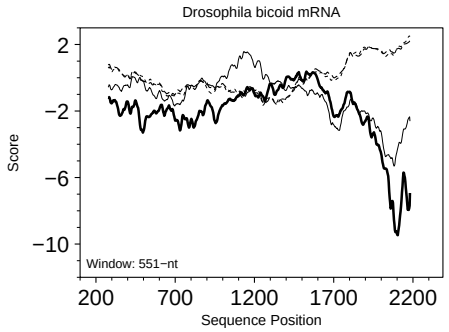

Figure 2: Distributions of the SIGSCR and STBSCR in *Drosophila bicoid* mRNA. The detected UFR is located from position 1814 to 2364 (or 175-725 in the 3'UTR). For further details see the caption to Figure 1.

Large number of sequences compiled in the UTRdb [13] are valuable for various statistical and structural analyses in order to understand the biological functions of 3'UTRs. Usually, 3'UTR is longer than 5'UTR and has a lower percentage of the GC composition in the sequence. In this study we found the unusual dsRNA in the 3'UTR from the database UTRdb. In Monte Carlo simulations, these large dsRNA structures are extremely significant in comparison with corresponding randomly shuffled sequences. Previous results from this laboratory have indicated that significantly UFRs are often closely correlated with functional structured RNA elements. The preliminary experimental data of cDNA encoding a cytotoxic ribonuclease has suggested that its 3'UTR contains a translational repression element [11] that may correlate with the unusual dsRNA structure (740 nts). The dsRNA structures found here are statistically significant and they are not be expected by chance.

Importance of the RNA duplex conformation is also observed in the other eukaryotic 3'UTR [10-14]. It has also reported that the individual gene in *C. elegans* can be specifically and potently suppressed by microinjection of a corresponding segment of dsRNA [7]. Inhibiting effects of exogenous dsRNA are post-transcriptional in RNA-based gene regulation. Although the mechanism for dsRNA mediated genetic interference is not well known, the exogenous dsRNA does cause early degradation of

homologous mRNAs [8]. It has also been suggested that the termination of transcription, lack of a poly(A) tail or failure to be translated may all make a mRNA aberrant [8]. The aberrant single-stranded RNA may convert into dsRNA, which could enter the RNA interference pathway by the ribonuclease Dicer [12]. In light of the recent evidence that dsRNA is a potent silencer of genes in animal and plant, it is intriguing to know if the dsRNA-mediated gene inhibition can be generated for organisms where genetic material cannot be delivered by microinjection. Does the endogenous dsRNA in the 3'UTR perform a similar interference in the gene inhibition under some conditions? Recent data of virus infection and expression assay indicate that the detected small dsRNAs in HIV-1 can induce antiviral RNAi in human cells [20]. We expect that the statistically significant dsRNAs detected in the database could provide useful information for further studies in exploring the RNA-based regulation of gene expression.

Acknowledgments

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

1. Liebhaber, S.A. (1997) mRNA stability and the control of gene expression. *Nucleic Acids Symposium Series* **36,** 29-32.

2. Belasco, J.G. and Braverman, G. eds. (1993) *Control of Messenger RNA Stability.* Academic Press Inc., San Diego, USA.

3. Bashirullah, A., Cooperstock, R.L. and Lipshitz, H.D. (1998) RNA localization in development. *Annu. Rev. Biochem.* **67,** 335-394.

4. Simons, R.W. and Grunberg-Manago M. eds. (1998) *RNA Structure and Function* Cold Spring Harbor Lab. Press, New York.

5. Hazelrigg, T. (1998) The destinies and destinations of RNAs. *Cell* **95,** 451-460.

6. Macdonald, P.M. and Kerr, K. (1998) Mutational analysis of an RNA recognition element that mediates localization of bicoid mRNA. *Mol Cell Biol* **18,** 3788-3795.

7. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature.* **391,** 806-811.

8. Zamore, P.D. (2002) Ancient pathways programmed by small RNAs. *Science* **296,** 1265-1269.

9. Dubnau, J. and Struhl, G. (1996) RNA recognition and translational regulation by a homeodomain protein. *Nature* **379,** 694-699.

10. Bergsten, S.E. and Gavis, E.R. (1999) Role for mRNA localization in translational activation but not spatial restriction of nanos RNA. *Development* **126,** 659-699.

11. Chen, S.-L., Le, S.-Y., Newton, D.L., Maizel, J.V.Jr. and Rybak, S.M. (2000) A gender-specific mRNA encoding a cytotoxic ribonuclease contains a 3' UTR f unusual length and structure. *Nucl. Acids Res.* **28,** 2375-2382.

12. Ambros, V. (2001) Dicing up RNAs. *Science* **293,** 811-812.

13. Pesole, G., Liuni, S., Grillo, G. and Saccone, C. (1998) UTRdb: a specialized database of 5'- and 3'-untranslated regions of eukaryotic mRNAs. *Nucl. Acids Res.* **26,** 192-195.

14. Le, S.-Y., Chen, J.-H. and Maizel, J.V.,Jr. (1990) *In:* Sarma, R.H. and Sarma, M.H. (eds.) Structure & Methods: Human Genome Initiative and DNA Recombination. Adenine Press, Schenectady, New York, Vol. 1, pp 127-136.

15. Le, S.Y., Chen, J.H., Konings, D., Maizel, J.V. Jr. (2003) Discovering well-ordered folding patterns in nucleotide sequences. *Bioinformatics* **19,** 354-461.

16. Laferriere, A., Gautheret, D. and Cedergren, R. (1994) An RNA pattern matching program with enhanced performance and portability. *CABIOS* **10,** 211-212.

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

18. Jaeger, J.A., Turner, D.H. and Zuker, M. (1989) Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA* **86,** 7706-7710.

19. Ferrandon, D., Elphick, L., Nusslein-Volhard, C. and St. Johnston, D. (1994) Staufen protein associates with the 3'UTR of bicoid mRNA to form particles that move in a microtubule-dependent manner. *Cell* **79,** 1221-1232.

20. Bennasser, Y., Le, S.-Y., Benkirane, M., and Jeang, K.-T. (2005) Evidence that HIV-1 encodes a siRNA and a suppressor of RNA silencing. *Immunity* . in press.

21. Le, S.-Y., Chen, Jih-H. and Maizel, J.V., Jr. (2003) Statistical inference for well-ordered structures in nucleotide sequences. *Proceedings of the 2003 IEEE Bioinformatics Conference, CSB 2003.* IEEE Computer Society Press, Los Alamitos, CA. pp. 190-196.

22. Cech, T.R., Tanner, N.K., Tinoco, I., Weir, B.R., Zuker, M. and Perlman, P.S. (1983) Secondary structure of the tetrahymena ribosomal RNA intervening sequence: *Proc. Natl. Acad. Sci. USA* **80,** 3903-3908.

Table 1. Distinct dsRNA structures found in the 3'UTRs.

mRNA Sequences	Accession Number	mRNA Length (bp)	3'UTR Length (bp)	RNA duplex in the 3'UTR Location in the 3'UTR	SIGSCR (mean and std)	STBSCR
In 7503 Human mRNAs						
pH-sensing regulator factor	AB001328	1704	1011	404-496/917-1010	-10.80 (-1.98,1.85)	-4.63
KIAA0432 gene	AB007892	5962	3710	2658-2796/3317-3450	-32.76 (-2.40,5.90)	-5.06
VHL gene	AF010238	4862	4148	2510-2821/2994-3307	-25.56 (-6.88,6.17)	-2.69
Aurora/IPL1-related kinase (AIK)	D84212	2033	655	99-426	-39.84 (-4.03,9.69)	-3.81
Alpha-fucosyltransferase	D87942	3088	1945	389-1169	-7.35 (-1.96,1.36)	-0.47
Adipogenesis inhibitory factor	X58377	2281	1618	810-1353	-8.07 (-2.73,1.75)	-1.40
Ring zinc-finger protein	U41315	3686	1886	568-1347	-12.25 (-2.24,2.20)	-1.52
Obese protein (ob)	U43653	3426	2866	501-571/1025-1095	-16.02 (-2.77,2.18)	-4.06
KIAA0349 gene	AB002347	6158	2330	729-796/952-1019	-7.00 (-0.97,1.40)	-2.04
GM2 activator protein	X62078	2436	1796	689-759/1197-1288	-6.59 (-2.00,1.71)	-2.08
KIAA0186 gene	D80008	3248	2563	1209-1344/1856-1995	-12.40 (-3.73,4.31)	-2.36
MACH-alpha-1 protein	X98172	2887	1156	263-323/658-718	-15.29 (-2.05,2.56)	-3.97
NAD+ dependent dehydrogenase	J05594	2412	1594	1260-1594	-31.08 (-1.75,3.03)	-6.34
SnoA protein	X15217	2875	918	174-300/645-773	-12.78 (-1.67,2.57)	-4.22
In 2457 other mammalian mRNAs						
Bovine ACTH receptor	X74501	2909	1890	295-436/829-979	-11.58 (-1.91,1.96)	-4.47
Bovine cyclase	U95958	5956	2317	837-1006/1290-1462	-17.66 (-1.99,2.79)	-4.48
Dog pinin	U77716	3893	1517	775-889/951-1066	-22.68 (-0.46,2.87)	-6.08
In 7633 rodent mRNAs						
Mouse NDPK-A	AF033377	887	388	203-387	-22.28 (-1.38,3.65)	-5.83
Mouse caspase-3	Y13086	1297	396	172-396	-32.83 (-1.76,6.04)	-5.36
Mouse rds protein	X14770	2632	1379	759-839/1014-1092	-13.32 (-1.18,1.92)	-5.20
Rat type IV collagenase	X71466	3231	769	416-523/659-766	-22.80 (-0.74,2.69)	-6.19
In 3499 other vertebrate mRNAs						
Goldfish aromatase	AB009335	2906	1287	162-306/517-671	-13.42 (-2.09,2.71)	-1.89
Zebrafish PAX7A	AF014367	2320	1140	499-753	-18.35 (-1.25,3.32)	-4.61
Goldfish CYPXIX	U18974	2939	1287	184-303/516-634	-21.13 (-2.04,3.51)	-3.13
Rainbow trout p53	M75145	2342	944	45-229/347-520	-14.47 (-1.75,3.09)	-2.33
Torpedo marmorate	U05591	2693	917	287-391/477-582	-16.12 (-1.17,4.00)	-4.41
X.laevis transcription factor A	U35728	2003	930	145-508	-29.37 (-3.66,6.94)	-3.46
X.laevis CTX	U43330	3135	2102	295-903	-30.98 (-3.92,6.35)	-3.36
X.laevis DNA polymerase	U49509	4396	581	89-230/389-533	-20.41 (-1.64,3.57)	-5.51
X.laevis myosin L-chain	Z33999	1305	759	140-309/482-653	-28.14 (-6.17,8.85)	-2.44
X.laevis tanabin	M99387	7019	1707	954-1091/1332-1466	-12.38 (-1.37,1.85)	-4.14
X.laevis receptor Xtll	U67886	3627	2004	676-929	-26.52 (-1.96,4.01)	-5.63
In 5067 invertebrate mRNAs						
F.lan hydroxylase	AJ001677	2870	1552	339-1093	-75.05 (-11.7,19.7)	-3.34
B.mori homeodomain	M64336	2874	1193	82-818	-31.27 (-4.91,9.92)	-3.70
B.mori lysozyme	L37416	1294	842	54-228	-15.98 (-1.09,2.70)	-3.97
C. elegans pha-1	X73845	2229	722	165-492	-14.15 (-2.03,2.73)	-4.99
C. elegans tra-1	M93256	4697	1205	151-459	-10.33 (-0.73,1.83)	-4.13
S. franciscanus bindin	M59490	2291	648	88-244/491-645	-35.11 (-1.89,5.88)	-4.57
Drosophila bicoid	X14458	2456	817	175-725	-10.28 (-2.08,1.93)	1.44
In 8116 plant mRNAs						
Fava bean ADP-glucose pyrophosphorylase	X76940	2040	424	1-328	-34.25 (-2.19,6.65)	-4.88
In 1154 fungi mRNAs						
Yeast virus 1, ScV1	X02232	819	745	1-133/394-520	-20.07 (-7.71,6.23)	-2.06