3580–3583 *Nucleic Acids Research, 1997, Vol. 25, No. 18*

Making (anti)sense of non-coding sequence conservation

David J. Lipman*

National Center for Biotechnology Information, National Library of Medicine, NIH, Building 38A 8N803, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received June 12, 1997; Revised and Accepted July 23, 1997

ABSTRACT

A substantial fraction of vertebrate mRNAs contain long conserved blocks in their untranslated regions as well as long blocks without silent changes in their protein coding regions. These conserved blocks are largely comprised of unique sequence within the genome, leaving us with an important puzzle regarding their function. A large body of experimental data shows that these regions are associated with regulation of mRNA stability. Combining this information with the rapidly accumulating data on endogenous antisense transcripts, we propose that the conserved sequences form long perfect duplexes with antisense transcripts. The formation of such duplexes may be essential for recognition by post-transcriptional regulatory systems. The conservation may then be explained by selection against the dominant negative effect of allelic divergence.

Since the early 1980s many studies on particular genes have noted sequence conservation in the 3′ untranslated regions (UTRs) of vertebrate mRNAs (1–3). Duret *et al.* (4) estimated that >30% of vertebrate mRNAs had conserved regions in their 3′ UTRs, defined as sharing at least 70% identity over >100 nucleotides between corresponding homologous genes (orthologs). They also noted the less frequent but still significant conservation in 5′ UTRs. We have recently observed long stretches of protein coding regions without silent changes in a substantial fraction of vertebrate mRNAs; most of these contain unusually conserved blocks both in the coding regions and in 5′ or 3′ UTRs (H.Sicotte and D.Lipman, unpublished data). A representative sample from a comparison of human and mouse orthologs is shown in Table 2. These conserved sequences are essentially unique in the genome and thus match only to corresponding regions of orthologous mRNAs in other species. The observed level of conservation is far greater than expected for non-coding regions or synonymous sites in coding regions on the basis of known evolutionary rates and divergence times (5).

What function constrains these regions? Sequence specific recognition, e.g., by RNA binding proteins, is an unlikely explanation because of the length of the conserved sequences. Furthermore, because so many different mRNAs contain these conserved regions, which are unique for each set of orthologs, sequence specific recognition would lead into an almost infinite regress. With >30% of the genes containing these unique conserved regions, then another 30% of the genes would be needed to code for these binding proteins, not to mention the proteins regulating these binding proteins, and so on. One might posit that many of these different sequences share common RNA secondary structure thus reducing the number of different binding proteins, but the sequence conservation would remain a mystery. It has been shown that short AU rich motifs promote mRNA degradation (6). Such motifs are often seen in the conserved portions of 3′ UTRs but these cannot explain the striking conservation between orthologs either. Another possibility would be that the conservation is due to the encoding of a protein on the complementary strand. Extensive database searches using translations of the complementary strand to these conserved regions did not reveal homologies to known proteins which could explain this conservation (results not shown).

A number of studies provide evidence that the conserved regions in 3′ UTRs are required for the regulation of mRNA stability (7). Typically deletion of these regions render the mRNA unresponsive to regulatory signals which normally lead to destabilization (8–10). Conversely, introduction of these regions into reporter mRNAs make them responsive to regulated destabilization (11–13). Conserved regions in 5′ UTRs (14) and coding regions (15–17) have also been implicated in regulation of mRNA stability.

The large number of bases in conserved blocks suggests a base-pairing interaction between mRNA and another nucleic acid. Over the last several years there has been an increasing number of reports of antisense RNA transcripts encoded by the complementary strand of a gene (18–22). Although most reported examples do not show evidence of coding regions, in some cases these countertranscripts encode expressed proteins (23,24). These countertranscripts are sometimes found in different tissues or developmental stages than their corresponding sense mRNA and thus a regulatory role for endogenous antisense has been proposed (25–28). Examples of regulation of gene expression by endogenous antisense have also been described for nematode (29), dictyostelium (30) and prokaryotes (31).

*Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: lipman@ncbi.nlm.nih.gov

Why would the antisense-based regulatory mechanism require sequence conservation? If cells have a destabilization/degradation system which specifically recognizes long, nearly perfect RNA duplex, then mutations in a region corresponding to a duplex will be selected against because of their mismatch with the other allele (Fig. 1). Consider, for example, the developmental expression pattern for Hoxa 11 sense and antisense transcripts (27); where sense transcripts are at high levels, antisense transcripts are at low levels, and vice-versa. When the Hoxa 11 antisense is abundant, most sense transcripts will be duplexed. Assuming the rate of transcription for the two alleles is roughly equal, a mutation in a region corresponding to a duplex would result in approximately half the sense transcripts forming mismatched duplexes. Let us further assume that the half life of a sense transcript is 12 h and the half life of a perfectly matching sense/antisense duplex is 12 min. When most of the sense transcripts are in perfect duplexes the drop in mRNA levels could therefore be an order of magnitude or more. However, a mutation leading to allelic divergence in a complementary region could lead to defective recognition of approximately half of the sense/antisense duplexes; thus, half the sense transcripts would have a half life of 12 min and half would have a half life approaching 12 h. The endogenous antisense mechanism would then only be able to reduce mRNA levels by a factor of two. Thus, the conserved regions in mRNAs will be maintained through selection against allelic divergence. In the three cases where the endogenous antisense has been sequenced and the corresponding orthologous mRNA sequences are also available, there is a strong correlation of complementary segments and sequence conservation. For example, in the BFGF gene, there is a single silent change between human and rat sequences in the 280 bases of the coding region which overlap the antisense transcript (unpublished observations).

With this hypothesis, one would predict that a chromosomal translocation in a region corresponding to a duplex would lead to upregulation of the product of the normal allele. An interesting example of this is the bcl-2/IgH translocation seen in B-cell lymphomas which is associated with increased levels of bcl-2 mRNA and bcl-2 protein as well as detectable levels of a bcl-2/IgH antisense transcript (32). Note that the translocation occurs within the 3′ UTR which contains a number of conserved blocks on either side of the breakpoint. Oligonucleotides complementary specifically to this chimeric antisense downregulate the bcl-2 gene product leading to apoptosis while oligonucleotides complementary to the bcl-2/IgH sense transcript have no effect (32,33). Presumably the chimeric antisense binds to the normal bcl-2 sense mRNA but is not efficiently recognized by the destabilization/degradation system and thus it acts as a competitive inhibitor of the normal bcl-2 antisense transcript.

Table 2. Examples of conserved blocks in human/mouse orthologous mRNAs[a]

mRNA	Accession no. (human)	Conserved regions (in nt) 5′ UTR length (%)	Coding region[b] identical blocks	3′ UTR length (%)
Immediate-early response protein NOT	X75918		153, 172, 150	
Human polyposis locus (DP2.5 gene)	M73548		147, 199	
Octamer binding transcription factor 1 (OTF1)	L20433		116, 125	
Homeobox protein hox-c4 (hox-3e) (cp19)	X07495		124, 136	
Acute phase response factor	L29277		178	69 (96%)
RNA binding protein EWS	X79233		133	156 (97%)
hnRNP-E2	X78136		209	167 (91%)
Eukaryotic initiation factor 4AII	D30655		167, 116	345 (96%)
				190 (96%)
Fibrillin	L13923		151	484 (87%)
Glutamate receptor 2 (HBGR2)	L20814	158 (85%)	258, 157	202 (98%)
p68 protein	X52104		175	301 (97%)
Thryoid hormone receptor α (c-erbA-1)	X55005	173 (91%)	183	80 (96%)
S-adenosylmethionine decarboxylase	M21154	122 (95%)	92, 139, 134	119 (88%)
			119 (97%)	
Sodium- and chloride-dependent taurine transporter	Z18956		160	69 (94%)
Transcription activator ZFX	X59739		159	575 (86%)
Homeobox c8 protein	M16938	208 (98%)	278	184 (85%)
				163 (88%)
Leukemia virus receptor 1 (GLVR1)	L20859	152 (94%)	145	145 (94%)
				171 (88%)
Very low density lipoprotein receptor	L20470		112	431 (93%)
Nervous system-specific octamer-binding transcription factor n-Oct 3	Z11933	60 (97%)	147	
Glutamate (NMDA) receptor subunit ξ1	D13515		139	78 (96%)
Voltage-dependent L-type Ca channel α1 subunit	Z34822		137	84 (99%)
				101 (95%)

[a]A representative sample of human mRNAs with long, perfect blocks of identity in the coding regions when aligned with the orthologous mouse mRNA. Conserved regions in 5′ and 3′ UTRs are also indicated when present.

[b]Using a synonymous mutation rate between mouse and man of 0.475 (W.Makalowski and M.S.Boguski, manuscript in preparation) a conservative bound on the probability of finding at least one identical block of length 75 in a coding sequence of 400 residues is ~3.8×10^{-5}.

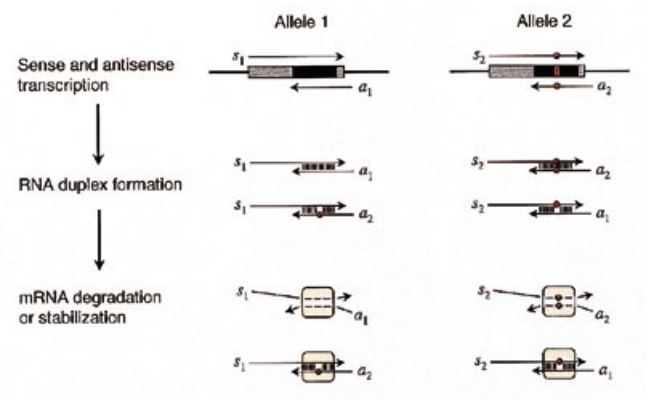

Figure 1. Endogenous antisense and sequence conservation. **Sense and antisense transcription**. s_1 is the sense transcript from Allele 1, a_1 is the antisense transcript from Allele 1, s_2 and a_2 are the transcripts from Allele 2. Sequence conservation is observed in the blackened region in both alleles but Allele 2 has a mutation shown in red. **RNA duplex formation**. The sense and antisense transcripts are complementary in the paired region (corresponding to the black bar in the above line). The s_1/a_1 and s_2/a_2 duplexes match perfectly but the s_1/a_2 and s_2/a_1 duplexes contain a mismatch. **mRNA degradation or stabilization**. The yellow box represents the destabilization/degradation protein(s). The perfect duplexes are efficiently processed (dashed and diagonal lines) while the mismatched duplexes remain intact.

Double-stranded RNA adenine deaminase (DSRAD) has been implicated in the destabilization of the BFGF mRNA base-paired with its countertranscript (34). The modification efficiency of DSRAD has also been shown to decrease exponentially as the length of RNA duplex drops below 100 bp (35). More recently, several additional DSRAD-related proteins have been sequenced (36,37), and other regulatory proteins specific for double-stranded RNA have also been characterized including the interferon-inducible protein kinase, PKR (38) and dsRNA-dependent RNAase L (39,40).

Whatever the proteins involved in destabilizing sense/antisense duplexes, if recognition is duplex-specific and not sequence specific one escapes the infinite regress trap. Perhaps even more importantly, strict specificity for near-perfect duplexes appears to be essential for the function of the postulated regulatory system, as otherwise recognition and degradation of other cellular RNA duplexes such as structural RNA would be catastrophic. The antiviral role of proteins recognizing long RNA duplexes (41,42) may be a serendipitous benefit of this regulatory system.

Studies on the regulation of the antisense transcripts of Wilms tumor suppressor (43), eif2-α (44) and myc (45) show that when the sense transcript is upregulated, the antisense transcription decreases, and when the sense is downregulated, the antisense transcription increases. Thus upregulating the gene increases the sense/antisense ratio and downregulating the gene decreases this ratio. If the sense/antisense duplexes are rapidly degraded a model with direct coupling of transcriptional regulation and mRNA stability appears straightforward.

For example, consider a gene where transcription of the sense message is 10-fold greater than the antisense transcription. In this context, the rapidly degraded duplexes are of little consequence. But decreasing sense transcription by a factor of only two or three and concomitantly increasing antisense transcription by the same magnitude would have a dramatic effect on overall mRNA stability and thus on mRNA levels. Such a model would help explain the drop in mRNA stability seen in a wide variety of systems including differentiation of MEL cells (46,47). For mRNAs with short half-lives, such as VEGF, the sense/antisense transcription ratio may be closer. The increase in VEGF mRNA levels with hypoxic induction is a result of a relatively small increase in transcription coupled with a significant increase in mRNA stability (48,49). Recent results by Kumar and Carmichael (50) show that polyoma virus sense/antisense duplexes, modified by DSRAD, are blocked from transport out of the nucleus, and this, rather than a drop in mRNA stability, accounts for the decreased levels of sense transcript. Their results suggest that the nucleus is the primary site of action for this proposed post-transcriptional regulation. Possible interference with the double-stranded RNA antiviral response provides additional support for this hypothesis. Whether stability, transport or both mechanisms are involved, the key for the model proposed here is that the specificity of recognition be conferred by long, near-perfect RNA duplexes.

Additional evidence for this model comes from experiments where treatment with oligonucleotides unexpectedly stabilized mRNA levels or upregulated a gene product. An oligonucleotide antisense to the start codon for myc (51) stabilized the mRNA level and blocked apoptosis while an analogous one for CD23 (52) increased the level of the CD23 gene product. Oligos in the sense orientation to the start codon (used as controls in experiments with antisense oligos) for the IGF-I receptor (53) and NF-κB (54) unexpectedly upregulated the respective gene products. In all four cases, the oligonucleotide was in a conserved region, suggesting that they, in some way, interrupted a duplex, thus inhibiting the endogenous destabilization/degradation system. These results suggest a simple approach for testing the model and perhaps modulating gene expression.

Other than coding for proteins or structural RNAs, the extensive ortholog-specific conservation in vertebrate mRNAs is perhaps the most pervasive functional constraint on the genome, as evidenced by sequence conservation. Any explanation for this conservation must deal with the problem of recognizing a unique signal for ~30 000 different mRNAs. The model of mRNA stability regulation by countertranscripts proposed here handles this infinite regress by positing recognition of nearly perfect, long duplexes, which depends not on a unique signal for each mRNA but still results in sequence conservation. The direct coupling of transcriptional regulation and post-transcriptional regulation of mRNA stability inherent in this model could be important in development, cellular differentiation, stress response, or any other situation of coordinated regulation of multiple genes. If correct, the mechanism proposed here may be modulated for therapeutic benefit.

ACKNOWLEDGEMENTS

I would like to thank S. Brenner for initial discussions on non-coding conservation; K. Katz, J. Hensold, A. Pause, R. Klausner, D. Botstein and R. Roberts for helpful discussions; A. Krainer for pointing out the potential effect of endogenous antisense on the induction of interferon and the antiviral response; G. Schuler, J. Wootton, R. Tatusov, S. Altschul, D. Landsman, for initial analyses of the data and discussion; and E. Koonin for all of the above and help with the manuscript. G. Schuler contributed the schematic.

REFERENCES

1 Hobart, P. M., Shen, L. P., Crawford, R., Pictet, R. L. and Rutter, W. J. (1980) *Science*, **210**, 1360–1363.
2 Ellison, J., Buxbaum, J. and Hood, L. (1981) *DNA*, **1**, 11–18.
3 Yaffe, D., Nudel, U., Mayer, Y. and Neuman, S. (1985) *Nucleic Acids Res.*, **13**, 3723–3737.
4 Duret, L., Dorkeld, F. and Gautier, C. (1993) *Nucleic Acids Res.*, **21**, 2315–2322.
5 Li, W. H., Luo, C. and Wu, C. (1985) In MacIntyre, R. J. (ed.) *Molecular Evolutionary Genetics*. Plenum Press, New York, pp. 1–94.
6 Zubiaga, A. M., Belasco, J. G. and Greenberg, M. E. (1995) *Mol. Cell. Biol.*, **15**, 2219–2230.
7 Ross, J. (1996) *Trends Genet.*, **12**, 171–175.
8 Ho, V., Acquaviva, A., Duh, E. and Bunn, H. F. (1995) *J. Biol. Chem.*, **270**, 10084–10090.
9 Tsai, K. C., Cansino, V. V., Kohn, D. T., Neve, R. L. and Perrone-Bizzozero, N. I. (1997) *J. Neurosci.*, **17**, 1950–1958.
10 Laird-Offringa, I. A., Elfferich, P. and van der Eb, A. J. (1991) *Nucleic Acids Res.*, **19**, 2387–2394.
11 McGowan, K. M., Police, S., Winslow, J. B. and Pekala, P. H. (1997) *J. Biol. Chem.*, **272**, 1331–1337.
12 Shaw, G. and Kamen, R. (1986) *Cell*, **46**, 659–667.
13 Lee, W. M., Lin, C. and Curran, T. (1988) *Mol. Cell. Biol.*, **8**, 5521–5527.
14 Roy, N., Laflamme, G. and Raymond, V. (1992) *Nucleic Acids Res.*, **20**, 5753–5762.
15 Wellington, C. L., Greenberg, M. E. and Belasco, J. G. (1993) *Mol. Cell. Biol.*, **13**, 5034–5042.
16 Wisdom, R. and Lee, W. (1991) *Genes Dev.*, **5**, 232–243.
17 Shetty, S., Kumar, A. and Idell, S. (1997) *Mol. Cell. Biol.*, **17**, 1075–1083.
18 Taylor, E. R., Seleiro, E. A. and Brickell, P. M. (1991) *J. Mol. Endocrinol.*, **7**, 145–154.
19 Rivkin, M., Rosen, K. M. and Villa-Komaroff, L. (1993) *Mol. Reprod. Dev.*, **35**, 394–397.
20 Campbell, C. E., Huang, A., Gurney, A. L., Kessler, P. M., Hewitt, J. A. and Williams, B. R. (1994) *Oncogene*, **9**, 583–595.
21 Silverman, T. A., Noguchi, M. and Safer, B. (1992) *J. Biol. Chem.*, **267**, 9738–9742.
22 McGuinness, T., Porteus, M. H., Smiga, S., Bulfone, A., Kingsley, C., Qiu, M., Liu, J. K., Long, J. E., Xu, D. and Rubenstein, J. L. (1996) *Genomics*, **35**, 473–485.
23 Volk, R., Koster, M., Poting, A., Hartmann, L. and Knochel, W. (1989) *EMBO J.*, **8**, 2983–2988.
24 Batshake, B. and Sundelin, J. (1996) *Biochem. Biophys. Res. Commun.*, **227**, 70–76.
25 Li, A. W., Seyoum, G., Shiu, R. P. and Murphy, P. R. (1996) *Mol. Cell. Endocrinol.*, **118**, 113–123.
26 Knee, R. S., Pitcher, S. E. and Murphy, P. R. (1994) *Biochem. Biophys. Res. Commun.*, **205**, 577–583.
27 Hsieh-Li, H. M., Witte, D. P., Weinstein, M., Branford, W., Li, H., Small, K. and Potter, S. S. (1995) *Development*, **121**, 1373–1385.
28 Murashov, A. K. and Wolgemuth, D. J. (1996) *Brain Res. Mol. Brain Res.*, **37**, 85–95.
29 Lee, R. C., Feinbaum, R. L. and Ambros, V. (1993) *Cell*, **75**, 843–854.
30 Hildebrandt, M. and Nellen, W. (1992) *Cell*, **69**, 197–204.
31 Wagner, E. G. and Simons, R. W. (1994) *Annu. Rev. Microbiol.*, **48**, 713–742.
32 Capaccioli, S., Quattrone, A., Schiavone, N., Calastretti, A., Copreni, E., Bevilacqua, A., Canti, G., Gong, L., Morelli, S. and Nicolin, A. (1996) *Oncogene*, **13**, 105–115.
33 Morelli, S., Delia, D., Capaccioli, S., Quattrone, A., Schiavone, N., Bevilacqua, A., Tomasini, S. and Nicolin, A. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 8150–8155.
34 Kimelman, D. and Kirschner, M. W. (1989) *Cell*, **59**, 687–696.
35 Nishikura, K., Yoo, C., Kim, U., Murray, J. M., Estes, P. A., Cash, F. E. and Liebhaber, S. A. (1991) *EMBO J.*, **10**, 3523–3532.
36 Melcher, T., Maas, S., Herb, A., Sprengel, R., Higuchi, M. and Seeburg, P. H. (1996) *J. Biol. Chem.*, **271**, 31795–31798.
37 Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P. H. and Higuchi, M. (1996) *Nature*, **379**, 460–464.
38 Proud, C. G. (1995) *Trends Biochem. Sci.*, **20**, 241–246.
39 Silverman, R. H. (1994) *J. Interferon Res.*, **14**, 101–104.
40 Dong, B., Xu, L., Zhou, A., Hassel, B. A., Lee, X., Torrence, P. F. and Silverman, R. H. (1994) *J. Biol. Chem.*, **269**, 14153–14158.
41 Jacobs, B. L. and Langland, J. O. (1996) *Virology*, **219**, 339–349.
42 Lee, S. B. and Esteban, M. (1993) *Virology*, **193**, 1037–1041.
43 Malik, K. T., Wallace, J. I., Ivins, S. M. and Brown, K. W. (1995) *Oncogene*, **11**, 1589–1595.
44 Noguchi, M., Miyamoto, S., Silverman, T. A. and Safer, B. (1994) *J. Biol. Chem.*, **269**, 29161–29167.
45 Chang, Y., Spicer, D. B. and Sonenshein, G. E. (1991) *Oncogene*, **6**, 1979–1982.
46 Mechti, N., Piechaczyk, M., Blanchard, J. M., Marty, L., Bonnieu, A., Jeanteur, P. and Lebleu, B. (1986) *Nucleic Acids Res.*, **14**, 9653–9666.
47 Hensold, J. O., Stratton, C. A., Barth, D. and Galson, D. L. (1996) *J. Biol. Chem.*, **271**, 3385–3391.
48 Levy, A. P., Levy, N. S., Wegner, S. and Goldberg, M. A. (1995) *J. Biol. Chem.*, **270**, 13333–13340.
49 Shima, D. T., Deutsch, U. and PA, D. A. (1995) *FEBS Lett.*, **370**, 203–208.
50 Kumar, M. and Carmichael, G. G. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 3542–3547.
51 Fischer, G., Kent, S. C., Joseph, L., Green, D. R. and Scott, D. W. (1994) *J. Exp. Med.*, **179**, 221–228.
52 Fournier, S., Rubio, M., Delespesse, G. and Sarfati, M. (1994) *Blood*, **84**, 1881–1886.
53 Delafontaine, P., Meng, X. P., Ku, L. and Du, J. (1995) *J. Biol. Chem.*, **270**, 14383–14388.
54 McIntyre, K. W., Lombard-Gillooly, K., Perez, J. R., Kunsch, C., Sarmiento, U. M., Larigan, J. D., Landreth, K. T. and Narayanan, R. (1993) *Antisense Res. Dev.*, **3**, 309–322.