A conserved sequence motif in 3′ untranslated regions of ribosomal protein mRNAs in nematodes

ASHWIN HAJARNAVIS and RICHARD DURBIN
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, United Kingdom

ABSTRACT

The 3′ untranslated regions (3′ UTR) of eukaryotic genes can contain motifs involved in regulation of gene expression or localization at the post-transcriptional level. This study concerns the identification of novel, conserved elements in 3′ UTRs of many ribosomal protein mRNAs in *Caenorhabditis elegans* and *Caenorhabditis briggsae.* Analysis of the region around the polyadenylation signal in many ribosomal protein mRNAs indicates the conservation of a sequence motif UUGUU occurring both before and immediately after the polyadenylation signal. Building a statistical model of this motif and searching a database of *C. elegans* 3′ UTRs reveals that this motif is also present in the 3′ UTR of some genes involved in translation and ribosome maturation, among others. We suggest that this signal may be involved in translation or other message-level regulation of ribosomal genes in *C. elegans.*

Keywords: ribosomal mRNA polyadenylation

INTRODUCTION

It has been observed that much of the regulation of synthesis of the translational apparatus is at the translational level (Meyuhas 2000). Ribosomal protein mRNAs in mammals and other organisms including frog and chicken commonly contain a 5′ terminal oligopyrimidine tract (TOP) (Levy et al. 1991), which is thought to bind to La protein (Cardinali et al. 1993) with cellular nucleic acid binding protein binding downstream (Pellizzoni et al. 1997). Subsequently, other genes involved in translation and its regulation have been found to have TOP mRNAs (Meyuhas 2000). The studies carried out in vertebrates suggest that there is a precedent for searching for some form of regulation at the mRNA level of ribosomal protein genes in invertebrates, and here we propose an element that may be involved in this process in the nematodes.

An important aspect of nematode molecular biology is the phenomenon of *trans*-splicing (Blumenthal 1995). Approximately 70% of *Caenorhabditis elegans* genes are *trans*-spliced, including all but two of the ribosomal proteins. The efficiency of the *trans*-splicing reaction and the proximity of the splices sites to the coding start means that these genes have a very short 5′ untranslated region (UTR), often of just a few bases. There are only two ribosomal protein genes that have long 5′ UTR sequences as determined by expressed sequence tag (EST) alignment (JC8.3a and W09C5.6b). A large number of the supporting ESTs for the former start with ACTTTT, which is pyrimidine rich and is potentially a TOP sequence. However, given the short 5′ UTRs in many nematode ribosomal protein mRNAs, it could be that elements involved in their common control are in the 3′ UTR.

RESULTS

While searching *C. elegans* and *Caenorhabditis briggsae* 3′ UTRs for conserved motifs, we noticed that the region around the polyadenylation signal of ribosomal protein genes appeared to be more conserved than expected.

After running MEME (Bailey and Elkan 1994) we identified a motif containing UUGUU adjacent to the polyadenylation signal. From 136 sequences (68 orthologous pairs) of ribosomal protein gene 3′ UTRs that have matching orthologs in *C. elegans* and *C. briggsae*, we found 57 that had UUGUU on both sides of the polyadenylation signal (Fig. 1A). In many of the other cases a UUGUU was present on one side, and in some cases there were multiple UUGUU motifs on one side, as in rows three and five of Figure 1A. The full alignment of 57 sequences with at least one UUGUU on each side is available as supplemental data (see http://www.sanger.ac.uk/Software/analysis/pajhmma/).

This alignment was used to build a specialized hidden Markov model. This model resembles the one used to

Reprint requests to: Richard Durbin, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; e-mail: richard.durbin@sanger.ac.uk; fax: +44-1223-494919.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.51306.

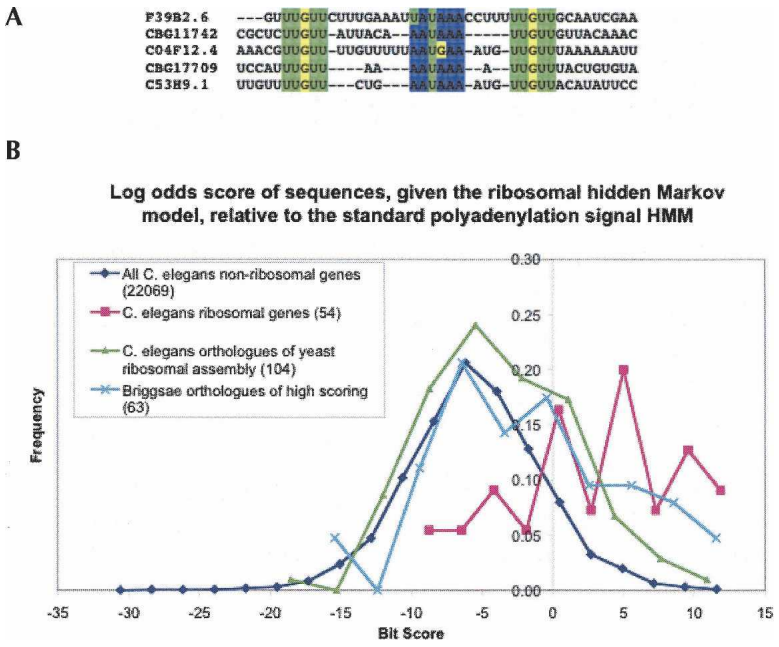

FIGURE 1. (*A*) A hand-edited alignment showing five of the 57 ribosomal protein 3′ UTR sequences from *C. elegans* and *C. briggsae*, containing UUGUU motifs on either side of the polyadenylation signal. (*B*) The bit score histogram resulting from finding the log(2) probability of various 3′ UTR sequence sets under the ribosomal model minus the log(2) probability under the standard model. (Dark blue) All *C. elegans* nonribosomal protein genes—this is the background distribution; (pink) *C. elegans* ribosomal protein genes; (green) *C. elegans* orthologs of yeast ribosomal assembly complex; (light blue) *C. briggsae* orthologs of *C. elegans* genes scoring >7.5 bits.

detect *C. elegans* polyadenylation signals, but has two extra components: weight matrices to model each of the two UUGUU motifs and distinctive length distributions modeling the lengths of the two motifs upstream and downstream of the polyadenylation signal. A score calculated as the difference between the log likelihood under this model and the standard polyadenylation signal model (Hajarnavis et al. 2004) can be used to determine how closely a given polyadenylation signal fits to the observed pattern. The distribution of these scores was calculated for sequences from four sets:

1. 22,069 *C. elegans* nonribosomal protein 3′ UTRs.
2. 54 *C. elegans* ribosomal protein 3′ UTR sequences, which were not included in model training.
3. 104 sequences of 3′ UTRs from *C. elegans*. The proteins of these genes represent the best BLASTP hit for 165 proteins from *Saccharomyces cerevisiae* that are implicated in pre-ribosomal complex formation in yeast (Fromont-Racine et al. 2003). The set includes only a few ribosomal proteins.
4. 63 *C. briggsae* orthologs of the 100 genes from set 1 that had the highest bit score under the ribosomal model.

Figure 1B shows that the score distributions for ribosomal and nonribosomal proteins are different ($p < 0.001$ in a Kolmogorov–Smirnov test). The peaks in the 0- and 5-bit regions are caused by single and double mismatches, respectively, to UUGUU, either upstream or downstream of the polyadenylation signal. The *C. elegans* orthologs of the yeast proteins involved in ribosome assembly also appear to have a different distribution from the nonribosomal proteins in general; there is a shoulder to the right suggesting that a subfraction of this protein set has higher scores and so may contain the motif. However this is not strong enough to be significant under a Kolmogorov–Smirnov test ($p = 0.08$).

The highest scoring 100 of the nonribosomal predictions (~0.5% of the total) score >7.5 bits. These 100 predictions come from genes that may therefore contain the motif, and we explored whether they might have some function related to that of the ribosomal protein genes. Most of these (77) have some annotation evidence, either from WormBase or from analysis of protein domains and BLASTP homologies to better annotated proteins. This set of 77 *C. elegans* high scoring predictions with annotation is available as supplemental data (see http://www.sanger.ac.uk/Software/analysis/pajhmma/). Eighteen of the 77 (23%) have functions related to translation. Genes in this annotated set include three genes related to eukaryotic translation factors, five involved in tRNA synthesis and processing, and 11 contributing to ribosomal and rRNA maturation. These can be seen in Table 1. It seems quite plausible that genes such as fibrillarin, which is involved in rRNA processing, should be under common control with the ribosomal protein genes. It is additionally promising that fibrillarin has the highest bit score in *C. briggsae*.

For those high-scoring *C. elegans* genes with a putative *C. briggsae* ortholog (63%), ribosomal motif log odds scores were also calculated for the ortholog's 3′ UTR. The distribution of these is also seen in Figure 1B and is significantly skewed to the right ($p = 0.001$ in a Kolmogorov–Smirnov test), suggesting that the motif tends to be conserved to some extent. However, the signal appears to be conserved across species in only a fraction of genes. To see whether the UUGUU motif is present in other species we analyzed the 40-bp sequence upstream of the poly-A tract from 625 mRNA sequences of non-*Caenorhabditis* ribosomal proteins in the EMBL data bank. There were 75 matches to UUGUU in total in these sequences (0.12 per sequence), compared to 204 matches in 108 sequences from *C. elegans* and *C. briggsae* (1.89 per sequence). The only

TABLE 1. A subset of the *C. elegans* genes having polyadenylation signals closest resembling those seen in ribosomal proteins

Elegans CDS	*Elegans* log odds score	*Briggsae* CDS	*Briggsae* log odds score	Homolog description
Y48A6B.3	10.909	CBG18231	10.620	NOLA2 Nucleolar protein family A; a core component of H/ACA small nucleolar ribonucleo-protein particles, which are involved in ribosome synthesis
F10E9.11	10.878	CBG16573	−3.314	Yeast FYV7 involved in processing 20S pre-rRNA; called CCDC59 in mammals (coiled coil protein)
F10E7.5	10.711	CBG13068	2.064	MRT4 involved in mRNA turnover and ribosome assembly; localizes to nucleolus
W06H3.2	10.681	CBG23897	−5.100	*pus-1*: PUS1 tRNA pseudouridine synthase
C28H8.11a	10.228	CBG09046	—	TDO2 Tryptophan 2,3-dioxygenase
ZK524.3b	9.65	CBG11879	—	*lrs-2*: LARS2 Leucyl tRNA synthetase, prob mitochondrial
T01C3.7	9.196	CBG11588	11.559	*fib-1*: FBL Fibrillarin–nucleolar rRNA processing
Y45F10D.7	9.079	CBG22378	3.040	Yeast UTP21 possible U3 snoRNP protein (mammalian WDR36)
Y56A3A.11	8.807	no_briggsae	—	SEN2 tRNA-splicing endonuclease subunit
K07E8.7	8.688	CBG19546	1.800	RPUSD2 (yeast RIB2) tRNA pseudouridine synthase
C01B10.8	8.577	CBG05389	4.274	KIAA0859 has S-adenosyl-methione dependent methyltransferase activity
W02A11.1	8.096	CBG13601	2.567	Yeast GCD14 tRNA (1-methyladenosine) methyltransferase subunit
Y24D9A.4c	7.995	CBG01675	7.699	RPL7A 60S ribosomal protein rpl-7a (yeast RPL8)
F18A11.6	7.758	CBG13174	—	SNAPC3 snRNA activating protein complex 50 kDa
T23D8.7	7.734	CBG03777	5.666	EIF2C-like; related to eukaryotic initiation factor 2C
F36A2.2	7.337	CBG12371	8.207	Yeast DUS1 dihydrouridine synthase modifies pre-tRNA(Phe)
C07E3.2	7.268	CBG02729	−4.740	NOC2 component of complex involved in intranuclear transport of ribosomal precursors
W04B5.4	7.148	CBG15659	−6.639	MRPL30 mitochondrial ribosomal protein L30

These have a log odds score that is within the top 0.5% of scores. These are the 18 (of 77) whose annotation suggests involvement in translation.

non-*Caenorhabditis* nematode ribosomal protein mRNA (from *Ascaris suum*) had two matches to UUGUU. In comparison, there were 70 matches to AACAA in non-*Caenorhabditis* sequences and only six in the *Caenorhabditis* sequences. When we analyzed the data by species, we again saw no comparable signal outside nematodes to that seen in *Caenorhabditis* (data not shown). It appears that the increased prevalence of the UUGUU motif around the polyadenylation signal is specific to nematodes.

DISCUSSION

The highest-scoring set of nonribosomal genes appears to contain a number of genes that could reasonably be expected to be co-regulated with ribosomal protein genes. However, the appearance in this set of some genes that are unlikely to be involved in translation suggests that the motif alone may not be specific for this function.

Bearing in mind the width of the distribution of the bit scores of ribosomal protein 3′ UTRs (Fig. 1B) and the observation that many ribosomal sequences were discarded from the 136 total during model building to arrive at 57, the function, if any, provided by this motif may be highly specialized within translation.

The UUGUU motif is reminiscent in its composition of the GU-rich sequence often present downstream of the cleavage site in vertebrate proteins (Zhao et al. 1999). We have previously reported that *C. elegans* mRNAs do not generally contain a GU-rich sequence downstream of the cleavage site (Hajarnavis et al. 2004). This region is thought to be important for the binding of Cleavage Stimulation Factor (CstF), which is part of the polyadenylation and cleavage apparatus (Zhao et al. 1999). Although it has been shown in *C. elegans* that the presence of CstF, rather than that of its binding site, is critical (Huang et al. 2001; MacDonald and Redondo 2002), it may be that the UUGUU motif reported here acts through binding or increasing the affinity for CstF, albeit upstream of its usual position in vertebrates.

In conclusion, we suggest the UUGUU motif may be involved either in 3′ end cleavage and polyadenylation of ribosomal proteins in nematodes or in regulation at the mRNA level, either of stability or translation. It would be possible to examine this experimentally.

MATERIALS AND METHODS

Sequences representing possible 3′ UTRs from 84 ribosomal proteins were extracted from WormBase (http://www.wormbase.org/).

The ribosomal motif information was incorporated into a generalized hidden Markov model (HMM) for the whole 3′ end region using the PAjHMMA framework (http://www.sanger.ac.uk/Software/analysis/pajhmma). The software used allows us to model sequences that have a distinctive length distribution by using a generalized HMM. The ribosomal polyadenylation signal HMM is derived from the standard polyadenylation signal model but has UUGUU motifs states inserted either side of the AAUAAA motif

states. The ribosomal model forces each sequence to pass through all the motifs. The UUGUU motifs themselves are built empirically, with probability 1/100 for a mismsatch and 97/100 for a match. In the third column, the occurrence of A is penalized to a slightly lesser degree than the others, scoring 5/100, since UUAUU motifs are occasionally seen where there is no UUGUU.

For HMM decoding, the forward algorithm (Durbin et al. 1998) was used to calculate $P(x)$, the probability of the sequence given the model for both the standard polyadenylation site model and the model with the motif added. The difference between the logs of these probabilities was used as a bit score (log likelihood ratio).

The Kolmogorov–Smirnov tests for nonparametrically assessing significance of the difference between two different distributions were performed at the Web site http://www.physics.csbsju.edu/stats/KS-test.

ACKNOWLEDGMENTS

The Wellcome Trust Sanger Institute is supported by The Wellcome Trust. A.H. was funded by the Medical Research Council.

Received February 4, 2006; accepted June 2, 2006.

REFERENCES

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (eds. R. Altman et al.), pp. 28–36. AAAI Press, Menlo Park.

Blumenthal, T. 1995. *Trans*-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet.* **11:** 132–136.

Cardinali, B., Di Cristina, M., and Pierandrei-Amaldi, P. 1993. Interaction of proteins with the mRNA for ribosomal protein L1 in. *Xenopus* Structural characterization of in vivo complexes and identification of proteins that bind in vitro to its 5′UTR. *Nucleic Acids Res.* **21:** 2301–2308.

Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis*. Cambridge University Press, Cambridge, UK.

Fromont-Racine, M., Senger, B., Saveanu, C., and Fasiolo, F. 2003. Ribosome assembly in eukaryotes. *Gene* **313:** 17–42.

Hajarnavis, A., Korf, I., and Durbin, R. 2004. A probabilistic model of 3′ end formation in *Caenorhabditis elegans*. *Nucleic Acids Res.* **32:** 3392–3399.

Huang, T., Kuersten, S., Deshpande, A.M., Spieth, J., MacMorris, M., and Blumenthal, T. 2001. Intercistronic region required for polycistronic pre-mRNA processing in *Caenorhabditis elegans*. *Mol. Cell. Biol.* **21:** 1111–1120.

Levy, S., Avni, D., Hariharan, N., Perry, R.P., and Meyuhas, O. 1991. Oligopyrimidine tract at the 5′ end of mammalian ribosomal protein mRNAs is required for their translational control. *Proc. Natl. Acad. Sci.* **88:** 3319–3323.

MacDonald, C.C. and Redondo, J.L. 2002. Reexamining the polyadenylation signal: Were we wrong about AAUAAA? *Mol. Cell. Endocrinol.* **190:** 1–8.

Meyuhas, O. 2000. Synthesis of the translational apparatus is regulated at the translational level. *Eur. J. Biochem.* **267:** 6321–6330.

Pellizzoni, L., Lotti, F., Maras, B., and Pierandrei-Amaldi, P. 1997. Cellular nucleic acid binding protein binds a conserved region of the 5′ UTR of *Xenopus laevis* ribosomal protein mRNAs. *J. Mol. Biol.* **267:** 264–275.

Zhao, J., Hyman, L., and Moore, C. 1999. Formation of mRNA 3′ ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.* **63:** 405–445.

A conserved sequence motif in 3′ untranslated regions of ribosomal protein mRNAs in nematodes

Ashwin Hajarnavis and Richard Durbin

RNA 2006 12: 1786-1789