Acta Protozool. (2005) 44: 231 - 236

ACTA
PROTOZOOLOGICA

The 3' Untranslated Region of mRNAs from the Ciliate *Nyctotherus ovalis*

Elodie DESTABLES[1], Nadine A. THOMAS[1], Brigitte BOXMA[2], Theo A. Van ALEN[2], Georg W. M. Van der STAAY[2], Johannes H. P. HACKSTEIN[2] and Neil R. McEWAN[1]

[1]Rowett Research Institute, Greenburn Road, Bucksburn, Aberdeen, Scotland; [2]Department of Evolutionary Microbiology, Faculty of. Science, Radboud University Nijmegen, Nijmegen, The Netherlands

Summary. The 3' untranslated regions (3'UTRs) of cDNAs from *Nyctotherus ovalis*, a ciliate from the digestive tract of cockroaches, were examined for their sequence composition. All 3' sequences studied here were characteristically short – generally having around 20 to 30 nucleotides between the stop codon and the first nucleotide of the polyA tail. The stop codon used in all sequences studied was UAA, which although one of the "universal" stop codons, is often used to encode glutamine in other ciliates such as *Tetrahymena*. The polyadenylation signal used in *N. ovalis* could not be determined from the sequence information, but it is clearly not the 'universal' AAUAAA signal. Furthermore, in messages encoding cathepsin B the 3'UTRs were of variable length, with the position where polyadenylation was initiated varying - despite a high conservation of the coding part of the message.

Key words: cathepsin, *Nyctotherus ovalis*, polyadenylation signals, stop codons.

INTRODUCTION

For any living organism, the proper control of expression of a gene is as important as the gene's informational content, i.e. the protein (or RNA) encoded by this particular gene. In its most simplistic form, a gene has an untranslated sequence located 5' of the coding sequence, and an untranslated sequence downstream of the coding sequence, representing the 3' end of the gene. The sequence 5' of the gene generally contains components such as promoters and enhancers, whereas the 3' sequence of the gene (3' UTR = Untranslated Region) is characterized by a number of motifs, which in eukaryotes, predominantly include sequences controlling the addition of a poly-A tail to the nascent mRNA chain (Zhao *et al.* 1999, Proudfoot and O'Sullivan 2002). Both the 5' UTR and 3' UTR cooperate in controlling the processing and translation of the mRNA (Proudfoot 2004).

Most mRNAs in eukaryotes possess a polyadenylate sequence, comprising anything from 30 to 200 adenylate residues at their 3' termini (Graber *et al.* 1999). This poly-A tail is added after transcription, in the course of the maturation of the mRNA, which occurs in the nucleus of the cell. First, the nascent RNA transcript is cleaved, and then the polyA sequence is added to the newly created 3' end of the message. In animals, cleavage occurs approximately 20 nucleotides downstream of a "polyadenylation signal", which, in the majority of the studies, is a hexanucleotide sequence

Address for correspondence: N. R. McEwan, Rowett Research Institute, Greenburn Road, Bucksburn, Aberdeen AB21 9SB, Scotland, U.K.; Fax. +44 1224 715349; E-mail: n.mcewan@rowett.ac.uk

(AAUAAA), which seemed to be highly conserved. Experimental studies revealed that a single substitution of any of these six nucleotides led to a significant decrease in the efficiency of both polyadenylation and RNA cleavage (Sheets *et al.* 1990, Graber *et al.* 1999).

Initially, the AAUAAA sequence (and its derivative AUUAAA - which can function at around 70% efficiency) was regarded as a universally conserved motif. More recently however, a number of studies have revealed that other eukaryotes make use of alternative polyadenylation motifs. Examples of organisms which use alternative motifs include yeast and plants, (Hyman *et al.* 1991, Rothnie 1996, Li and Hunt 1997, Graber 2003). Moreover, there can be a degree of flexibility which is not just restricted to sequence composition. For example there can be variation in the polyadenylation site based on the carbon sources used in experiments with yeast (Sparks and Dieckmann 1998) or with the cell-type (male germ cells) in mouse (MacDonald and Redondo 2002).

In general, studies analysing polyadenylation in organisms other than man, mice, *Drosophila*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Oryza sativa* are sparce. In particular, there is a remarkable lack of information about polyadenylation in protists, which represent a much greater biodiversity and divergence that animals, fungi and plants altogether. With respect to ciliates, which represent a particularly species-rich and very diverse taxon (Corliss 2004), only a few papers have been published dealing with the analysis of polyadenylation of a few genes (Williams and Herrick 1991, Liu and Gorovsky 1993, Ghosh *et al.* 1994, McEwan *et al.* 2000), and these papers deal with sequence analysis for proposed polyadenylation rather than providing experimental evidence. In the rumen ciliate *Entodinium caudatum* (McEwan *et al.* 2000), it was concluded that the "universal" AAUAAA signal was absent, although a similar sequence, AUAAA was often present in the region where the polyadenylation signal would normally be found. Similarly, the "universal" signal was not found in cDNAs from *Euplotes* (Ghosh *et al.* 1994), where the motif (A/U)UAAAA was proposed as a possible alternative polyadenylation signal. Limited sequence information led to the suggestion that *Oxytricha fallax* (Williams and Herrick 1991), uses three motifs-TAAAC, TGAAC and AGAAC, which might be described by the consensus sequence (tRAAC). In *Tetrahymena thermophila*, the motifs $TGTGT(N)_{1-8}TAA(N)_{0-11}AAGTATT$ have been described in four histone mRNAs (Liu and Gorovsky 1993). Moreover, the mating pheromone *Er*-1 in the ciliate *Euplotes raikovi* has been shown to have the unusual polyadenylation signal AACAAA (Miceli *et al.* 1992). Obviously, these consensus motifs have little sequence identity in common with each other, or with the "universal" polyadenylation signal thought to be characteristic of animals.

Here we describe a bioinformatical analysis of the 3' untranslated regions (UTR) of a number of cDNA sequences obtained by random sequencing of a cDNA library from the anaerobic ciliate *Nyctotherus ovalis* var. *Blaberus* sp. Amsterdam (Armophorea, Clevelandellida). This is an organism, which in keeping with other ciliate species, has two nuclei - a micronucleus (involved in sexual reproduction) and a macronucleus (involved in asexual reproduction). Within the macronucleus chromosomes have been shown to exist as highly amplified gene-sized mini-chromosomes (Akhmanova *et al.* 1998). Thus, every gene is located on a short chromosome with telomeres at both ends, and these genes function as a single transcription unit. Here we demonstrate that many, if not most, of these genes lack the "universal" polyadenylation signals.

MATERIALS AND METHODS

Obtaining cDNA clones. *Nyctotherus ovalis* from strain *Blaberus* spec. Amsterdam thrives in the hindgut of the giant cockroach (Van Hoek *et al.* 1998), where it has been maintained in laboratory culture for more than 12 years. Ciliates were harvested by electro-migration making use of their unique anodic galvanotactic swimming behaviour (Van Hoek *et al.* 1999). Immediately after isolation, the ciliates were lysed in 8 M guanidinium chloride. RNA was extracted with the aid of the Rneasy Plant mini-kit (Qiagen) following the recommendatiosn of the manufacturer. Total RNA was reverse transcribed, amplified, and the cDNA molecules cloned into pDNR-LIB (Clontech), and grown in *E. coli* strain DH5α by Genterprise, Mainz, Germany. Briefly, this was carried out as follows. The cDNA library was constructed using the "Creator SMART cDNA Library Construction" Kits from BD Biosciences/Clontech. cDNA was amplified using the primers CDS III and SMART IV (CDS III: 5´- ATT CTA GAG GCC GAG GCG GCC GAC ATG d(T)30N_1N 3´; SMART IV 5´- AAG CAG TGG TAT CAA CGC AGA GTG GCC ATT ACG GCC GGG - 3´). After amplification, the cDNAs were size fractionated on Sepahryl S500 columns, and cloned into pDNR-LIB.

DNA sequencing. 96 clones were picked at random and sequenced using the ABI Prism™ BigDye terminator sequencing ready reaction kit (Perkin Elmer Corporation, Norwalk, CT, USA) on an ABI Prism™ 377XL DNA sequencer (Perkin Elmer Corporation).

Only sequences where a stop codon could be identified unequivocally (based on a combination of Blast analysis and translation in three different reading frames), were included into the study. Se-

quences analysed further in this study have been deposited in the EBI sequence database with accession numbers AJ965632-AJ965670.

RESULTS AND DISCUSSION

In the current dataset 39 cDNA sequences were recovered, which allowed the unequivocal identification of a stop codon. Of these the 3' UTR region of 25 cDNAs were analysed initially, with 14 clones which encode cathepsin (cysteine protease) analysed separately in more detail (see below). Firstly, all of these clones used TAA as a stop codon, as observed earlier (Akhmanova *et al.* 1998, Voncken *et al.* 2002, Van Hoek *et al.* unpublished). This supports the hypothesis that *N. ovalis* uses a TAA or TAG to encode a stop codon (Knight *et al.* 2001, Lozupone *et al.* 2001), in contrast to other ciliates, which use TAA or TAG to encode glutamine (Hoffman *et al.* 1995, Lozupone *et al.* 2001). The absence of codons for TGA, which is a stop codon but in *Euplotes octocarinatus* is the only codon for cysteine (Grimm *et al.* 1998) and has previously been discussed as having an unknown function in *N. ovalis* (Knight *et al.* 2001, Lozupone *et al.* 2001), may argue for *N. ovalis* using the universal genetic code.

However, the sequences which have been studied previously for stop codon usage represent macronuclear genomic sequences. Hence it was unknown where the polyadenylation site would lie within the messages transcribed from these genes. What was clear from these sequences is that the "universal" polyadenylation signals could not be identified. Consequently, it remained unclear how much of these sequences is transcribed beyond the stop codon.

The sequences studied here were recovered from total RNA, and all are polyadenylated. Since, in addition, they are likely to be derived from a macronuclear minichromosome suggesting that they are likely to encode a functional gene. Fig. 1 shows, that the sequences between the stop codon and the poly-A tail are rather short for all of the cDNA sequences studied here. These sequences generally lack the motifs which have previously been described as polyadenylation signals in other organisms. Furthermore, there is no obvious alternative motif common to the 3' UTR of all of these sequences. This conclusion is based on visual inspection, attempts to

Fig 1. The sequence of the 3' UTR of 25 randomly selected cDNA clones from the ciliate *Nyctotherus ovalis*.

Sequence Nr	Gene Encoded	Stop codon	3'UTR Sequence
255	Hypothetical	TAA	TCTCTTAACACACAACGTTGAGTAAT
257	Lactoylglutathione lyase	TAA	GTAATTCCTCACTCTCATTTCCATTCCATG
259	Hypothetical	TAA	CCGCTTCTAACCCTTAACCCACTTATTCAC
260	Hypothetical	TAA	CCATAATTCACAGTTCC
261	Glutamate dehydrogenase	TAA	GCATACTAACCCATTCACTTATTGAC
262	Beta/alpha-amylase	TAA	ATGATTTATTACCGTCTACTGAATTC
277	Hypothetical	TAA	TCTCTTAACACACAACGTTGAGTAAT
279	Hypothetical	TAA	CCCATCTAATCTATCACCACACCCTGCTTTACAAG
284	Hypothetical	TAA	TTCACCTGCGCCGCACCAGTATGTCTTCAC
285	Ribosomal protein L32	TAA	TCCCTGCCATCTCACGTACATATCACACCACATGC
294	Hypothetical	TAA	CGATTAACAATATCTTTGAC
299	Hypothetical	TAA	TCTCTTAACACACAACGTTGAGTAACCT
303	Hypothetical	TAA	CTGCCTAACCCTGACTAACTCACGTGCTCCG
309	Cathepsin H	TAA	GCGACTAACGCTAACCTAACTTCTC
312	Ribosomal protein L21	TAA	GCTATTAACCTGCATACACCCAAGCATCCC
321	Hypothetical	TAA	AACCTTAATGGCTTAACTCACTTATTGATTTC
324	Beta tubulin	TAA	GACTTTAACATTCACAAAACCTTTACG
325	Hypothetical	TAA	GCCCTTAACCCTCATTAATTAATCC
335	Beta tubulin	TAA	GACTTTAACATTCACAAAACCTAG
496	Hypothetical	TAA	GCCCTTAACCCTGATTAATTAATGC
503	Hypothetical	TAA	ATTTCATTAACCCTCCGCTCACTCTTTCGATCC
508	Hypothetical	TAA	AACCTTAATGGCTTAACTCACTTATTGATTC
556	Hypothetical	TAA	GCTCTAACCCTCTCCTAATTAATTCAC
557	Hypothetical	TAA	AACCTTAATGGCTTAACTCACTTATTGATTCC
586	Hypothetical	TAA	CCGCTTCTAACCCTTAACCCACTTATTCGTG

Fig 2. Alignment of the DNA for 14 cDNA sequences from *Nyctotherus ovalis*. The stop codons are shown in bold and are underlined. Within the translated region of the gene only 3 nucleotides are different across all of these sequences. Two of these gives synonymous amino acid substitutions, the other (blocked in grey) gives rise to an amino acid change. Positions of identity to the top sequence are indicated '.'. To maintain alignment of the first 'A' of the poly(A) tail and the last nucleotide which is not 'A', the areas with no nucleotide are shown '-'.

```
311  ATGAACTACAAGAGCGGAGTTTTCAAGTGCCACAACACTAACTACATCGGAGGCCACGCCGTTCTCGCTATGGGATACCATGAGGAAGATGAGAAGGGAAAGAAGGATCCTAACTACGAAGTAAAGAAC
263  ....................................................................................................................................
330  ....................................................................................................................................
267  ....................................................................................................................................
290  ....................................................................................................................................
271  ....................................................................................................................................
329  ....................................................................................................................................
286  ....................................................................................................................................
337  ....................................................................................................................................
287  ....................................................................................................................................
316  ....................................................................................................................................
323  ....................................................G...............................................................................
328  ....................................................................................................................................
326  ....................................................................................................................................
     ************************************************************************************ **********************************

311  TCCTGGGGAGCCCACTGGGGACTTGCTGGATACTTCAGAATTGCACCAGGAACCTGCAATATGCAAGGAGGCGTTGTTTGCACAGAATTTTAAGACACTTAAGCAACTAATGCTAATCCAACCTG--
263  ..................................................................................CTAA..............................----
330  ..................................................................................CTAA..............................C---
267  ........G.........................................................................TAA..............................TAG
290  ..................................................................................CTAA..............................C.T-
271  ..................................................................................TAA..............................AC--
329  ..................................................................................TAA..............................AG---
286  ..................................................................................TAA..............................CTG-
337  ..................................................................................CTAA..............................C---
287  ..................................................................................TAA..........T.....................G---
316  ..................................................................................TAA..............................TGT
323  ..................................................................................TAA..............................G.--
328  ..................................................................................TAA..............................AC--
326  ..................................................................................TAA..............................AC----
     ******** ************************************************************************** ************* *****************
```

align sequences and searches carried out using the UTResource database (http://bighost.area.ba.cnr.it/BIG/UTRHome).

Notably, within the clones which were selected at random for sequencing were 14 which encoded a cysteine protease - cathepsin (Fig. 2). This number of copies of this cDNA may at first appear rather high. However, due to the variability seen immediately prior to the polyadenylation site (discussed in more detail below) it appears unlikely that this is a reflection solely due to cDNA bank amplification. Furthermore, cysteine proteases are known to have a major role to play in ciliates (e.g. Parama *et al.* 2004), suggesting that they may be represented as abundant messages.

In the cathepsin cDNAs the sequence immediately prior to the stop codon was analysed to evaluate levels of variability within this region. This allowed analysis back 219 nucleotides (encoding 73 residues) prior to the stop codon for each clone. In total, only 6 nucleotide substitutions were detected - 4 of which were identical (i.e. 6 from 3066 nucleotides) and with one exception these encoded synonymous substitutions - with the four identical substitutions occurring in the codon immediately prior to the stop codon. This level of sequence conservation is also observed in the first 28 nucleotides following the stop codon, with only one cDNA having a single nucleotide deviation from the others. However, in the subsequent nucleotides, which precede the polyA tail - ranging in length from 2 to 6 nucleotides - there are 12 unique sequences within the 14 clones (clones 330 and 337 are identical to each other, as are clones 271 and 328). This is a high level of variability, both in terms of sequence length and sequence content, and is unprecedented in the sequence prior to the stop codon.

The reason for this variability immediately prior to the polyadenylation site is unknown. A number of different mechanisms may be proposed as the answer to this dilemma, but there are probably three most likely causes: (1) it may be due to non-specific addition of nucleotides to the nascent message prior to polyadenylation; (2) the polymorphisms may have arisen in the DNA as a consequence of re-arrangements during the formation of macronuclear mini-chromosomes; or (3) a number of different copies of the gene exist within the micronucleus - effectively different paralogues which still share the same function due to conservation within the coding region. It is unclear which of these mechanisms is taking place, and they should not be considered as being mutually exclusive, with more than one of these factors being possible.

However, it is interesting to note that in a study which included analysis of different cDNAs from messages for actin genes (Ghosh *et al.* 1994) a similar observation was made whereby the actin messages were polyadenylated at one of two different sites. The authors point out that it is unclear if the two forms of messages co-exist in a cell, or if they are used in a developmental-specific manner. The range in number of different messages identified here, although not precluding developmental-specific versions of the message, would appear to argue against there only being a single form of the message present in different developmental stages.

Thus, it has to be concluded from the genes analysed here that: (i) *Nyctotherus ovalis* does not make use of any of the "universal" polyadenylation signals, or of putative polyadenylation signals which have been described previously; (ii) that there is no evidence for any obvious candidate sequence to signal the onset of polyadenylation; and (iii) that, at least in the case of the cathepsin cDNA clones, the precise initiation site for polyadenylation is also not clearly determined, since this region exhibits a hyper-variability in the region 29-34 nucleotides following the stop codon - both in terms of sequence length and composition.

Acknowledgements. This work was performed as part of the EU Framework V Quality of Life and Management of Living Resources Research Programme grant CIMES (QLK3-2002-02151). The Rowett Research Institute receives funding from SEERAD.

REFERENCES

Akhmanova A., Voncken F., Van Alen T., Van Hoek A., Boxma B., Vogels G., Veenhuis M., Hackstein J. H. P. (1998) A hydrogenosome with a genome. *Nature* **396:** 527-528

Corliss J. O. (2004) Why the world needs protists! *J. Euk. Microbiol.* **51:** 8-22

Ghosh S., Jaraczewski J. W., Klobutcher L. A., Jahn C. L. (1994) Characterization of transcription initiation, translation initiation, and poly(A) addition sites in the gene-sized macronuclear DNA molecules of *Euplotes*. *Nucleic Acids Res.* **22:** 214-221

Graber J. H. (2003) Variations in yeast 3'-processing cis-elements correlate with transcript stability. *Trends Genet.* **19:** 473-476

Graber J. H., Cantor C. R., Mohr S. C., Smith T. F. (1999) *In silico* detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc. Natl. Acad. Sci. USA* **96:** 14055-14060

Grimm M., Brunen-Nieweler C., Junker V., Heckmann K., Beier H. (1998) The hypotrichous ciliate *Euplotes octocarinatus* has only one type of $tRNA^{Cys}$ with GCA anticodon encoded on a single macronuclear DNA molecule. *Nucleic Acids Res.* **26:** 4557-4565

Hoffman D. C., Anderson R. C., DuBois M. L., Prescott D. M. (1995) Macronuclear gene-sized molecules of hypotrichs. *Nucleic Acids Res.* **23:** 1279-1283

Hyman L. E., Seiler S. H., Whoriskey J., Moore C. L. (1991) Point mutations upstream of the yeast ADH2 poly(A) site significantly reduce the efficiency of the 3'-end formation. *Mol. Cellular Biol.* **11:** 2004-2012

Knight R. D, Freeland S. J., Landweber L. F. (2001) Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet*. **2:** 49-58

Li Q., Hunt A. G. (1997) The polyadenylation of RNA in plants. *Plant Physiol*. **115:** 321-325

Lozupone C. A., Knight R. D., Landweber L.F. (2001) The molecular basis of nuclear genetic code change in ciliates. *Curr. Biol*. **11:** 65-74

Liu X., Gorovsky M. A. (1993) Mapping the 5' and 3' ends of *Tetrahymena thermophila* mRNAs using RNA ligase mediated amplification of cDNA ends (RLM-RACE). *Nucleic Acids Res.* **21:** 4954-4960

MacDonald C.C., Redondo J.L. (2001) Reexamining the polyadenylation signal: were we wrong about AAUAAA? *Mol. Cell Endocrinol*. **90:** 1-8

McEwan N. R., Eschenlauer S. C. P., Calza R .E., Wallace R. J., Newbold C. J. (2000) The 3' untranslated region of messages in the rumen protozoan *Entodinium caudatum*. *Protist* **151:** 139-146

Miceli C., La Terza A., Bradshaw R. A., Luporini P. (1992) Identification and structural characterization of a cDNA clone encoding a membrane-bound form of the polypeptide pheromone Er-1 in the ciliate protozoan *Euplotes raikovi*. *Proc. Natl. Acad. Sci USA.* **89:** 1988-1992

Parama A., Iglesias R., Alvarez M. F., Leiro J., Ubeira F. M., Sanmartin M. L. (2004) Cysteine proteinase activities in the fish pathogen *Philasterides dicentrarchi* (Ciliophora: Scuticociliatida). *Parasitology* **128:** 541-548

Proudfoot N. (2004) New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr. Opin. Cell. Biol*. **16:** 272-278

Proudfoot N., O'Sullivan J. (2002) Polyadenylation: a tail of two complexes. *Curr. Biol*. **12:** R855-R857

Rothnie H. M. (1996) Plant mRNA 3'-end formation. *Plant Mol. Biol.* **32:** 43-61

Sheets M. D, Ogg S. C., Wickens M. P. (1990) Point mutations in AAUAAA and the poly(A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*. *Nucleic Acids Res.* **18:** 5799-5805

Sparks K. A., Dieckmann C. L. (1998) Regulation of poly(A) site chice of several yeast mRNAs. *Nucleic Acids Res.* **20:** 4676-4687

Van Hoek A. H., van Alen T. A., Sprakel V. S., Hackstein J. H. P., Vogels G. D. (1998) Evolution of anaerobic ciliates from the gastrointestinal tract: phylogenetic analysis of the ribosomal repeat from *Nyctotherus ovalis* and its relatives. *Mol. Biol. Evol*. **15:** 1195-1206

Van Hoek A. H., Sprakel V. S., Van Alen T. A., Theuvenet A. P., Vogels G. D., Hackstein J. H. P. (1999) Voltage-dependent reversal of anodic galvanotaxis in *Nyctotherus ovalis*. *J. Euk. Microbiol*. **46:** 427-433

Voncken F., Boxma B., Tjaden J., Akhmanova A., Huynen M., Verbeek F., Tielens A.G., Haferkamp I., Neuhaus H. E., Vogels G., Veenhuis M., Hackstein J. H. P. (2002) Multiple origins of hydrogenosomes: functional and phylogenetic evidence from the ADP/ATP carrier of the anaerobic chytrid *Neocallimastix* sp. *Mol. Microbiol.* **44:** 1441-1454

Williams K.R., Herrick G. (1991) Expression of the gene encoded by a family of macronuclear chromosomes generated by alternative DNA processing in *Oxytricha fallax*. *Nucleic Acids Res.* **19:** 4717-4724

Zhao J., Hyman L., 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:** 405-45

Received on 21st February, 2005; revised version on 6th May, 2005; accepted on 12th May, 2005