1678–1684 *Nucleic Acids Research, 1997, Vol. 25, No. 9* © 1997 Oxford University Press

Ataxia-telangiectasia: structural diversity of untranslated sequences suggests complex post-transcriptional regulation of *ATM* gene expression

Kinneret Savitsky, Matthias Platzer[1], Tamar Uziel, Shlomit Gilad, Adam Sartiel, Andre Rosenthal[1], Orna Elroy-Stein[2], Yosef Shiloh and Galit Rotman*

Department of Human Genetics, Sackler School of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel, [1]Department of Genome Analysis, Institute of Molecular Biotechnology, Jena D-07745, Germany and [2]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel

Received February 17, 1997; Revised and Accepted March 21, 1997

DDBJ/EMBL/GenBank accession nos U67092, U67093

ABSTRACT

Mutations in the *ATM* gene are responsible for the multisystem disorder ataxia-telangiectasia, characterized by neurodegeneration, immune deficiency and cancer predisposition. While no alternative splicing was identified within the coding region, the first four exons of the *ATM* gene, which fall within the 5′ untranslated region (UTR), undergo extensive alternative splicing. We identified 12 different 5′ UTRs that show considerable diversity in length and sequence contents. These mRNA leaders, which range from 150 to 884 nucleotides (nt), are expected to form variable secondary structures and contain different numbers of AUG codons. The longest 5′ UTR contains a total of 18 AUGs upstream of the translation start site. The 3′ UTR of 3590 nt is contained within a single 3′ exon. Alternative polyadenylation results in 3′ UTRs of varying lengths. These structural features suggest that *ATM* expression might be subject to complex post-transcriptional regulation, enabling rapid modulation of ATM protein level in response to environmental stimuli or alterations in cellular physiological states.

INTRODUCTION

Ataxia-telangiectasia (A-T) is a multisystem autosomal recessive disorder characterized by cerebellar ataxia, immune deficiencies, increased cancer predisposition, chromosomal instability and radiation sensitivity (see refs 1,2 for recent reviews). The gene responsible for A-T was recently cloned by Savitsky *et al.* (3,4) using a positional cloning approach, and designated *ATM*. The *ATM* gene product, a protein of 3056 amino acids, is a member of a growing family of large proteins in various species that are involved in genome stability, cellular responses to DNA damage and cell cycle control. These proteins share a highly conserved carboxy-terminal region of ~300 amino acids showing high sequence homology to the catalytic domain of PI-3 kinases (4; reviewed in refs 5–7).

The *ATM* gene has 66 exons spread over 150 kb of genomic DNA (8). A band originally estimated at ~12 kb was observed by northern blot analysis in every tissue tested to date, and a minor 10.5 kb transcript was present in fibroblast RNA (3). Extensive RT–PCR analysis of fibroblasts and lymphoblasts did not reveal alternative splicing within the coding sequence (unpublished data). Since the open reading frame (ORF) of the *ATM* transcript consists of 9168 nucleotides (nt) (4), we predicted that at least 2.8 kb of the mature *ATM* transcript would consist of untranslated sequences.

An additional gene, designated *E14* by Byrd *et al.* (9,10) and *NPAT* (nuclear protein mapped to the A-T locus) by Imai *et al.* (11), is located immediately upstream of the *ATM* gene and is transcribed in the opposite direction, with only ~500 bp separating the transcription start sites of the two genes. *NPAT* encodes a putative serine/threonine-rich protein of 1427 amino acids. Sequences matching nuclear localization signals in the carboxyl terminus suggest this protein may be transported into the nucleus (11). The region upstream of both genes is embedded in a CpG island which covers the first exon and part of the first intron of each gene (9–11; M. Platzer *et al.*, submitted). This region contains a 'TATA-less' promoter, which might regulate bidirectional transcription of these divergently transcribed genes (10).

We characterized the 5′ and 3′ untranslated regions (UTRs) of *ATM* transcripts in several tissues and compared their sequence to that of the corresponding genomic domains of the gene. We observed extensive alternative splicing in the 5′ UTR and differential polyadenylation in the 3′ UTR. The structural diversity of these UTRs suggests complex regulation of the *ATM* gene.

*To whom correspondence should be addressed. Tel: +972 3 6408584; Fax: +972 3 6407471; Email: grotman@ccsg.tau.ac.il

MATERIALS AND METHODS

RNA isolation

Total RNA was extracted from a human primary fibroblast line, F-89, a lymphoblastoid cell line, L1000 (both derived from Caucasian males) and Jurkat T cells, using the Tri-Reagent system (Molecular Research Center, Inc., Cincinnati, OH). Poly(A)$^+$ RNA was isolated from total RNA using Dynabeads mRNA Direct Kit (Dynal, Oslo, Norway). Poly(A)$^+$ RNA from adult human cerebellum, thymus, testis and placenta was purchased from Clontech (Palo Alto, CA).

RACE analysis of 5′ and 3′ ends of *ATM* transcripts

5′ and 3′ RACE (rapid amplification of cDNA ends) analysis was performed on poly(A)$^+$ RNA from human cerebellum, testis, placenta and lymphoblasts using the Marathon cDNA Amplification Kit (Clontech). Adaptor-ligated double-stranded cDNA libraries were prepared essentially as suggested by the manufacturer. Superscript II Reverse Transcriptase (Gibco-BRL, Gaithersburg, MD) was used for the first strand synthesis. First round PCR was performed on these libraries for 30 cycles, using the Expand Long Template PCR System (Boehringer-Mannheim, Germany). A nested PCR approach was used to isolate 5′ and 3′ RACE products.

For 5′ RACE, first round PCR was carried out with an adaptor-specific primer and a gene-specific antisense primer (5′-GATGAATCTTTCACTGTATCCATG-3′) derived from nt 395–372 on the ORF. Nested PCR was then performed on these reactions for 20 cycles using an antisense primer (5′-GGTTCAC-AATTTCAGAACACACATC-3′) derived from sequences immediately upstream of the initiator codon at positions 30–6 on exon 4. For 3′ RACE, first round PCR was performed with a sense primer (5′-TCAGTGTTGGTGGACAAGTG-3′) derived from the end of the ORF (positions 9077–9096). Two additional primers were derived from the 3′ UTR (accession no. U67093) and used for separate nested PCR reactions: 5′-TGCATCATTTT-TCAGATCTCT-3′, from nt 844–864; and 5′-CTAAGATAGAA-AACTGCCAAGGAC-3′, from nt 2831–2854. All PCR reactions were carried out in the presence of 1.5 µg Flash Anti-*Taq* Monoclonal Antibody (Molecular Biology Resources, WI). In order to test their authenticity, RACE products were hybridized to a panel of cosmids spanning the *ATM* gene (3,4), before being subjected to sequence analysis.

Analysis by RT–PCR of the 5′ UTRs of *ATM* transcripts

First strand synthesis was carried out essentially as described previously (12), using either 5 µg of total RNA isolated from fibroblasts, lymphoblasts or Jurkat cells, or 1 µg of poly(A)$^+$ RNA from adult human cerebellum, thymus, testis and placenta (Clontech). The reactions, performed in a final volume of 20 µl, also contained Superscript II Reverse Transcriptase (Gibco, BRL), 1× buffer supplied by the manufacturer, 30 U of Rnasin (Promega, Madison, WI), and 10 pmol of a gene-specific antisense primer (5′-GACAGCCAAAGTCTTGAGGAAG-3′) from nt 735–714 in the *ATM* ORF.

RT–PCR for the identification of the various 5′ UTRs was carried out with primers derived from nt 46–65 on exon 1a (5′-GCGA-GAGGAGTCGGGATCTG-3′) or from nt 18–39 on exon 1b (5′-CTGAATTGAACCCTGCCTCCTG-3′) and an antisense primer (5′-GTTGACGGCAGCAGATAAGCAGATC-3′) derived from exon 4 between nt 43 and 19 of the ORF. One-twentieth of the reverse transcription products served as template, and the reactions were subjected to 30 amplification cycles. The PCR products were separated on agarose gels, and each individual product was further purified by touching the specific band on the gel with a toothpick, and reamplifying for 20 cycles.

Semi-quantitative RT–PCR analysis of the 5′ UTRs was performed by autoradiography of labeled PCR products, using the primers described above, which derive from exons 1a, 1b and 4. The relative amounts of total *ATM* transcripts were estimated by amplifying the same RT products using primers from nt 217–239 (5′-GAATGTCTGAGAATAGCAAAACC-3′) and 610–589 (5′-CGTCAGTCTGAGAACAGCATCC-3′) of the ORF. The number of cycles was optimized for each primer set to ensure that the amplifications did not reach a plateau. All PCR reactions were initially carried out for five cycles at an annealing temperature equal to the T_m of the primers, with additional cycles at annealing temperatures 4°C below the T_m (20 cycles for reactions with the primers from exons 1a or 1b and 4, and 16 cycles for those with primers from the ORF). The PCR reactions contained 2 µCi of [α-^{32}P]dCTP. The PCR products were electrophoresed on agarose gels, blotted onto Qiabrane plus nylon membranes (Qiagen, Hilden, Germany), and exposed to X-ray films for ~4 h.

Sequencing of PCR products

PCR products were prepared for sequencing using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany). DNA sequencing was performed using the dye terminator chemistry and the Ampli-*Taq* FS enzyme (Perkin Elmer). The products were analyzed on an automated DNA sequencer (Applied Biosystems, model 373A). Sequence contigs were assembled using the AutoAssembler program (Applied Biosystems Inc., Perkin-Elmer).

Computer analysis

Pairwise comparisons were done using the BESTFIT program in the sequence analysis software package developed by the Genetics Computer Group (GCG) at the University of Wisconsin, or the Global Alignment Program (GAP) (13). Prediction of secondary structures and calculation of free energy values of the 5′ UTRs were performed with the MFOLD program in the GCG software package.

RESULTS

Structural diversity of the 5′ UTRs of *ATM* transcripts

5′-RACE was carried out in order to map the transcriptional start sites of *ATM* transcripts. RT–PCR analysis was subsequently performed to further characterize the 5′ UTRs of these transcripts in various tissues and cell types. The sequence of two cDNA clones, Col-g and En-a (4), as well as that of RACE and RT–PCR products obtained from different tissues, was matched against the corresponding genomic sequence (accession no. U67092) obtained by sequencing a cosmid contig spanning the entire *ATM* gene (M. Platzer *et al*., submitted). The results indicated that there are four non-coding exons (leader exons 1a and 1b, and exons 2 and 3; Fig. 1) that undergo extensive alternative splicing, forming divergent 5′ UTRs (Fig. 2). Attempts to locate the transcriptional start sites of the *ATM* gene by primer extension resulted in multiple bands (data not shown), which are probably due to the diversity of the 5′ UTRs and/or their stable secondary structure.

```
gttgcttctt cctccagaag tgggcgctgg gcagtcacgc agggtttgaa
ccggaagcgg gagtaggtAG CTGCGTGGCT AACGGAGAAA AGAAGCCGTG
GCCACGGGAG GAGGCGAGAG GAGTCGGGAT CTGCGCTGCA GCCACCGCCG   Exon 1a
CGGTTGATAC TACTTTGACC TTCCGAGTGC AGTGgtaggg gcgcggaggc
aacgcagcgg cttctgcgct gggaaattca gtcgtgtgcg acccagtctg
tcctctcccc agaccgccaa tctcatgcac ccctccagag tggcccttga
ctcctccctc tcctcactcc atctttcctg gcctctctcc gggtgcttag
cggacttggc caataaccTC CTCCTTTTAA ACGCCCTGAA TTGAACCCTG
                     1b-1
CCTCCTGCGC ATCCTCTTTT TGTGTCACCT TAGGGTTCAG ATTTAACTAC
                     1b-2
GCGACTTGAC TAGTCATCTT TTGATCTCTC TCTCGTATTT AGTACTTTTA
GTCAGCGAGC ATTTATTGAT ATTTCAACTT CAGCCTCGCG GTTAAGAGCT
TGGGCTCTGG AATCATACGG CTGGAATTGG AATTCTGTCA GTCGTGTGGC
CGCTCTCTAC TGTCTTGTGA AGATAAGTGA GATAATCTTG ACCTGTGGTG
                     1b-3
AGCACTCGTG AGCGTTAGCT GCTGTATTTA CCAGGTACAG ATAAGACAAC   Exon 1b
TACAGTGGAT GATAATGTAT GTGGTGATAG GGGAGTACTC TGATGGTAGA
GGAGTGACTT TGGTTCTCTG CAAACTCAGC CTGAGACTAT CAATTCAGTT
TGTGGTGAGA CCTCGCAGTG TTACCTTGGC AGATGGTAGA AGCCTTCCAG
ATGGAAGGAA AAATGCGTGT AAAGGCACAA AGTGTAGAAG GACCCTGAAG
CTCCAGCGTG AGGCCTGGCA TTGAATGAAA TATATTTTGT GGGTTTTCAG
CTGCTGAAGT CATAGGAATG GATGAGACCA AGAAAACAAA GCTGTTTTTG
AGgtatgagc ggaagaagag atatcaggag actttcgaaa cagtcataac
ggaagttaat atgatcattg ctaacatttg ctgtgtttca gGCACTGTAA   Exon 2
GCATGTATAT GGGTCCTTAA AGGGACTCAT AGAGgtaggt actagtattg
tttttccttt tatcattgag aaactgaggt ttgaagagat tagtgaactt
.....................................................
tgagatatat gttccctgac agtttatcta aaatgattct ctctacagtt
attttctatt actgtgtttt tgtttcctca gAGGCATACA TCACAATTTG
GAATTATGCA TTGGTTTATC AATTTACTTG TTTATTGTCA CCCTGCTGCC   Exon 3
CAGATATGAC TTCATGAGGg taggatttgt atctgtttag ttcattattt
gtctagctat aagtagtaaa tattgtttgc atctatcagt gaatgagcat
.....................................................
aataattata tacacatttt ttcacacctc tttctctcta tatatgcata
tatacatata catatatata cctatatgta tttttttac agACAGTGAT
GTGTGTTCTG AAATTGTGAA CCATGAGTCT AGTACTTAAT GATCTGCTTA   Exon 4
TCTGCTGCCG TCAACTAGAA CATGATAGAG CTACAGAACG AAAGgtagta
aattacttaa attcaatttt tccttgaaat gtgtgattag taacccatta
```

Figure 1. Nucleotide sequence of the genomic region spanning the 5′ end of the *ATM* gene. Exons (in capital letters) are boxed. ATG codons are underlined and the ATG initiator codon of the major ORF is doubly underlined. Bent arrows denote transcriptional start sites determined by 5′ RACE. The 5′ TOP element at the 5′ end of exon 1b and 1b-2 is underlined with dashed lines. The inverted triangle indicates the 3′ end of exons 1b-1, 1b-2 and 1b-3. Only a partial sequence of the introns between exons 2, 3 and 4 is shown. The 5′ boundary of exon 1a replaces the one previously published in ref. 8, and is 4 nt downstream of it. The complete sequence of this region has been submitted to GenBank under accession no. U67092.

5′ RACE and RT–PCR identified two leader exons, 1a and 1b, and three shorter variants of 1b: exons 1b-1, 1b-2 and 1b-3. Exon 1a is particularly GC-rich (62%) and is part of the CpG island. It is located upstream of exon 1b, and separated from it by 184 bp (Fig. 1). The 634 bp long leader exon 1b appears in processed 5′ UTRs as several variant forms (Figs 1 and 2). The 5′ end is determined by three transcription initiation sites at nt 1 (exons 1b/1b-1), 48 (exon 1b-2) and 252 (exon 1b-3) of exon 1b. The 3′ ends of exon variants 1b-1, 1b-2 and 1b-3 are generated by the usage of an alternative donor site at nt 317–318 of exon 1b. Ten AUG codons are present in the distal 318 bp that are unique to exon 1b, in contrast to the AUG-less exons 1b-1, 1b-2 and 1b-3 (Figs 1 and 2). The 5′ region of exons 1b/1b-1 is pyrimidine-rich (73% in the proximal 53 nt), and contains a 5′ TOP (terminal oligopyrimidine) element of 10 pyrimidines. Exon 1b-2 begins with a tract of six pyrimidines. Such 5′ TOP elements have been shown to play a critical role in the mechanism that regulates the translation of ribosomal protein mRNAs and other cellular transcripts in a growth-dependent manner (reviewed in ref. 14). Interestingly, the short exon 1b-3 is devoid of these 5′ TOP elements and of the rest of the pyrimidine-rich sequences (Figs 1 and 2).

An acceptor splice site was not evident immediately upstream of exon 1b. Consistent with this, exon 1a was never found together with exon 1b or its derivatives in the same transcript, further supporting our assumption that these exons actually serve as alternative leader exons.

These leader exons can be joined to exon 4, which contains the translation initiator codon, either directly or via exons 2 and/or 3, harboring two and three upstream AUGs, respectively (Figs 1 and 2). In certain 5′ UTRs, the small intron of 89 bp between exon 1b and exon 2 is not spliced out, thereby adding two additional AUGs to the mature transcript (Figs 1 and 2). However, the possibility that transcripts containing this small intron are actually splicing intermediates cannot be ruled out.

The multitude of 5′ UTRs resulting from multiple transcription start sites and alternative splicing events are depicted in Figure 2. These mRNA leaders show considerable length variation, ranging from 146 to 884 nt, and are predicted to form extensive secondary structures with free energy values from –27.4 to –174.2 kcal/mol.

A total of 18 AUG codons are present in the longest 5′ UTR, upstream of the putative translation start site in exon 4 (Fig. 2). The consensus sequence for initiation of translation proposed by Kozak (15) is (A/G)CCAUGG. The two most critical positions for the function of this site are highly conserved: the large majority of vertebrate mRNAs have a purine (usually A) in position –3, and a purine (usually G) in position +4 (where the A of the AUG is designated +1) (15,16). The context of the initiation codon of the *ATM* ORF, ACCAUGA matches this consensus fairly well, whereas the contexts of most of the upstream AUGs have lower similarities to the consensus (Fig. 1). The possibility that some of these AUG codons are used to generate isoforms of the ATM protein is unlikely since there are stop codons in all reading frames downstream of these AUGs. However, translation initiation using any of these AUGs would result in small upstream ORFs (uORFs) ranging in length from 3 to 71 codons.

Abundance of different 5′ UTRs in various tissues

In order to analyze the relative abundance and tissue specificity of the different 5′ UTRs, we performed semi-quantitative RT–PCR (see Materials and Methods) on several tissues and cell types using an antisense primer located immediately downstream of the initiator codon in exon 4, and a sense primer derived from the 5′ end of either exon 1b or exon 1a. The PCR reactions were carried out under cycling conditions that did not allow the products to reach a plateau. These reactions were performed in the presence of a radioactive nucleotide, since under these conditions the products could not be detected by ethidium bromide staining. As a relative measure of the total amount of *ATM* transcripts, PCR was also performed on the same RT products using primers derived within the ORF. The results (Fig. 3) show that most of the

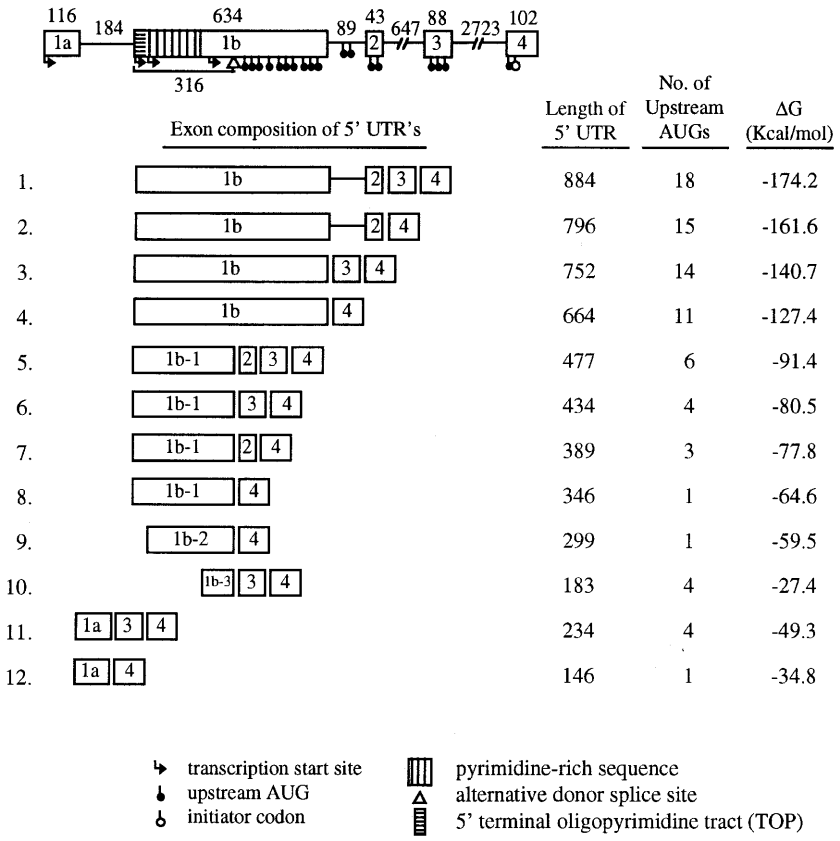

Figure 2. Structural diversity in the 5′ UTR of *ATM* transcripts. The top scheme shows the genomic organization of the 5′ region of the *ATM* gene. Exons are represented by boxes and introns by solid lines, with their sizes in bp shown above them. Exon composition and characteristics of 12 *ATM* 5′ UTRs are shown below.

5′ UTRs described in Figure 2 are present in all cell types and tissues tested, with some minor differences in the relative abundance of specific 5′ UTRs. Certain 5′ UTRs, such as those containing exon 1b-1 spliced to exon 2, were not detected under these PCR conditions, suggesting that these mRNA leaders might be more rare. The largest PCR product of ~1.6 kb contains the 647 bp intron, between exons 2 and 3. It is not clear whether such a 5′ UTR is present in mature *ATM* transcripts, or this product actually represents a splicing intermediate.

The 3′ UTRs of *ATM* transcripts

The previously published 3′ UTR of the *ATM* transcript was 538 bp long, and ended with a poly(A) tail (3; accession no. U26455). Nine *ATM* cDNA clones, isolated from oligo(dT)-primed cDNA libraries representing different tissues, contained the same 3′ UTR and poly(A) tail. However, no polyadenylation signal was found in this sequence. Comparison of this 3′ UTR to the genomic sequence derived from cosmid E3 (3,4; accession no. U82828) revealed that the terminal poly(A) tail is genomic, and belongs to the *Alu* element that occupies most of this 3′ UTR (3, and unpublished data). Attempts to identify additional downstream sequences in *ATM* transcripts using 3′ RACE with primers located within this UTR resulted only in products ending with this poly(A) tail, most likely because of preferred oligo(dT) priming on this stretch of 18 'A's (Fig. 4; positions 522–539 in accession no. U67093).

A 3.0 kb cDNA fragment, 3–6, previously identified in the process of isolating candidate cDNAs for A-T, was located to the same cosmid clone (E3) as the 3′ end of the *ATM* transcript. Genomic PCR using primers derived from the 3′ UTR and from clone 3–6 showed that this cDNA clone lies immediately downstream of the *Alu* element of the *ATM* transcript (Fig. 4). Sequence comparison of clone 3–6 with these genomic PCR products and with cosmid E3 revealed a perfect match. We reasoned that clone 3–6 might represent the continuation of the *ATM* 3′ UTR, and RT–PCR experiments supported this assumption (not shown). The longest *ATM* 3′ UTR is therefore 3590 bp long (accession no. U67093) and its poly(A) tail is preceded by a cannonical polyadenylation signal, AATAAA (Fig. 4). This UTR is contiguous with the stop codon of the ORF and is located within the last and largest exon of 3.8 kb (8). Two *Alu* elements were identified between nt 319–539 and 1488–1778 of this 3′ UTR, as well as several regions with a high content of A and U residues, and eight copies of the pentanucleotide AUUUA (Fig. 4). Such AU-rich elements (AREs) may have a role in destabilization of the mRNA (reviewed in refs 17,18).

3′ RACE analysis using a primer located downstream of the first *Alu* element resulted in three products of 1.1, 1.6 and 1.9 kb, obtained from lymphoblast, cerebellar and placental RNA, respectively. Their sequences indicated that they are contiguous and overlap with clone 3-6 (Fig. 4). Pairs of unusual polyadenylation signals were found within 10–30 nt upstream of the poly(A) tails of two 3′ RACE products (Fig. 4). Two AGTAAA variants precede the 1.6 kb product, whereas two GATAAA precede the 1.1 kb

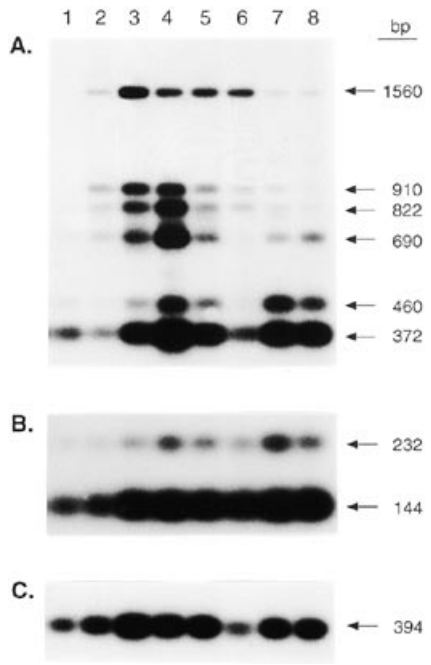

Figure 3. Autoradiogram of radioactively labeled RT–PCR products, obtained using RNA from the following sources: 1, lymphoblasts; 2, testis; 3, thymus; 4, cerebellum; 5, placenta; 6, Jurkat cells; 7, exponentially growing fibroblasts; 8, quiescent fibroblasts. Semi-quantitative RT–PCR was carried out using primers derived from exon 1b and 4 (**A**), exon 1a and 4 (**B**) and the ORF (**C**), as explained in Materials and Methods. The sizes of the PCR products differ from the length of the corresponding 5′ UTRs shown in Figure 2, according to the location of the primers used. The PCR products of 910, 822, 690, 460, 372, 232 and 144 bp correspond to 5′ UTRs nos 1, 2, 4, 6, 8, 11 and 12, respectively.

product. The poly(A) of the 1.9 kb product was found to be genomic (positions 2780–2796 in accession no. U67093), and is probably a result of oligo(dT) priming on this stretch of 17 ‘A’s. Several additional variant potential polyadenylation signals are found within this UTR, including two canonical AATAAA. One of these precedes the poly(A) tail of an additional 3′ RACE product obtained with a primer located downstream of the stretch of 17 ‘A’s (Fig. 4). These results indicate that different 3′ UTRs formed by alternative polyadenylation might be present in various *ATM* transcripts.

DISCUSSION

An extensive collection of *ATM* transcripts with diverse 5′ UTRs is produced by a combination of multiple transcription start sites and alternative splicing of non-coding 5′ exons. These divergent mRNA leaders, which differ markedly in length, in predicted secondary structure and in the number of upstream AUGs, might play an important role in the translational regulation of the *ATM* gene.

Gene expression can be controlled by regulating translation, thus allowing the cell to respond to environmental changes more rapidly than by changing the rate of transcription. Translation of certain mRNAs can be regulated by modulating either its rate of initiation, or its efficiency (19,20). Translational initiation is a multistep pathway aimed at positioning the ribosome at the appropriate initiating AUG codon. Traditionally, the cap-dependent pathway involves binding of the translation initiating complex to the cap structure of the mRNA, followed by scanning for an AUG in a favorable context for initiation (15,16,21). The length and sequence of the mRNA leader establish the intrinsic rate of initiation of translation, since secondary structures upstream of the AUG codon can block ribosome scanning (16,22). Upstream AUGs may also impair translation, as expected from the conventional scanning model. These AUGs might result in small uORFs, which have been shown to affect the translation of the major ORF (23).

The large majority of eukaryotic mRNAs have a short 5′ UTR of 20–100 nt with no upstream AUGs, enabling efficient translation. Five to ten percent of vertebrate mRNAs have long,

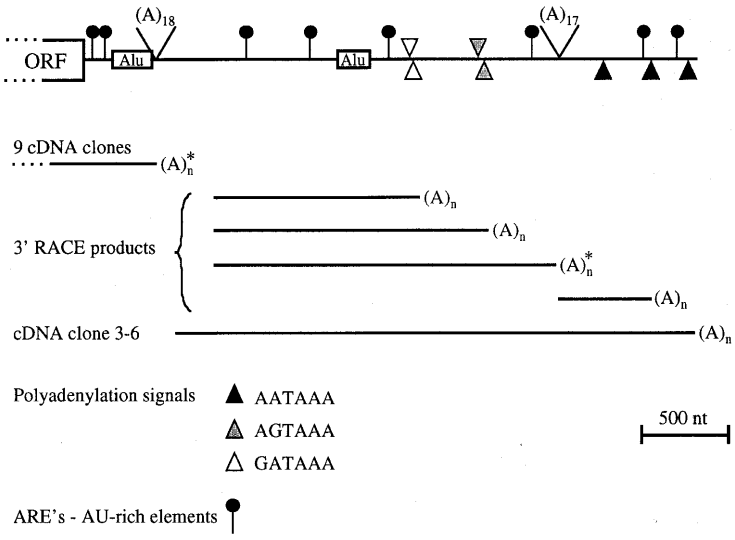

Figure 4. The 3′ UTR of the *ATM* gene. The top scheme shows the longest 3′ UTR with prominent sequence elements. The asterisks denote poly(A) tails of cDNA clones or RACE products which resulted from oligo(dT) priming on genomic stretches of ‘A’s. Below are shown the relative locations of 3′ RACE products and cDNA clones, indicating the existence of various *ATM* 3′ UTRs. The sequence of this region has been submitted to GenBank under accession no. U67093.

highly structured 5′ UTRs that usually contain upstream AUG codons. Such transcripts primarily encode regulatory proteins such as growth factors, transcription factors, proto-oncogenes, cytokine receptors and proteins involved in signal transduction and the immune response (16,21,24). The inhibitory features present in long 5′ UTRs can actually serve as regulatory elements, modulating the translatability of the respective transcript. Upstream ORFs have been shown to participate in selective translational regulation of specific mRNAs, providing in some instances control of tissue specific and developmentally regulated gene expression (23). Transcripts containing long, structured and uORF-burdened 5′ UTRs might be translated under specific physiological conditions due to the modification or temporary appearance of *trans*-factors capable of facilitating the melting of stable secondary structures. Alternatively, such factors might confer IRES (internal ribosome entry site) activity to these 5′ UTRs. Such elements serve to directly position the ribosome in the vicinity of the translation start codon, thereby enabling the bypass of upstream ORFs or secondary structures which are inhibitory for ribosomal scanning (recently reviewed in ref. 25). The 5′ UTR of the PDGF2 gene, for example, has recently been shown to contain an IRES element which is activated upon megakaryocytic differentiation of K562 cells (26).

The secondary structures of the long 5′ UTRs of the *ATM* transcripts have calculated free energy values from –27.4 to –174.2 kcal/mol. Some of these leaders should be potent inhibitors of translation, since secondary structures with a stability of ~–50 kcal/mol in the 5′ UTR were shown to inhibit translation by 85–95% (22). It is noteworthy that 10 out of 18 upstream AUGs are contained in the long leader exon 1b, and can be spliced out by usage of an alternative donor splice site (Fig. 2). The extensive alternative splicing that takes place among the non-coding exons considerably modulates the number of AUGs, so *ATM* transcripts may contain between 1 and 18 upstream AUGs in the different 5′ UTRs identified in this study. Thus, some of the transcripts are heavily burdened with uORFs, while others are almost devoid of these translation inhibitors.

Genes encoding transcripts with long, structured 5′ UTRs often use alternative promoters and/or splice sites to generate supplementary transcripts with short leader sequences devoid of translation inhibitory features (21,27). Such transcripts, containing short 5′ UTRs, most likely provide a constitutive basal level of gene expression, whereas the characteristic features of the longer 5′ UTRs provide extensive post-transcriptional regulation. An additional putative promoter region was identified within the fourth intron of the *ATM* gene, immediately upstream of the first coding exon. This region contains a ‘TATA-box’, and consensus binding sequences for several *trans*-acting factors (M. Platzer *et al.*, submitted). We could not prove the existence of a transcript driven by this promoter, although some of the shorter bands observed by primer extension might derive from it. Our inability to detect 5′ RACE products that originate from this promoter might be due to a lower transcription rate, or to transcription restricted to specific physiological conditions or cell types. As in other genes harbouring long cumbersome 5′ UTRs, the second *ATM* promoter might drive the expression of a transcript with a considerably shorter mRNA leader. Such a transcript should have higher translatability, being devoid of most uORFs, secondary structures and other inhibitory elements, and could provide a constitutive basal level of ATM protein synthesis.

The 3′ UTRs of eukaryotic mRNAs are also involved in post-transcriptional regulation. They contain signals that determine masking, stability and localization of the mRNA, as well as the efficiency of its translation (see refs 28,29 for reviews). The regulation of mRNA decay is an important mechanism for controlling the level of gene expression (17,30). In mammalian cells, mRNA half-lives can vary from minutes to days. Recent studies suggest that the stability of an mRNA is often suited to its function. Among the short-lived mammalian mRNAs are those that encode proteins that are expressed transiently in response to extracellular stimuli, such as growth factors, cytokines and transcription factors. AREs are found within the 3′ UTRs of a large number of labile mammalian mRNAs. Such sequences have been shown to confer mRNA instability (17,18). Secondary structures in the 3′ UTR can also bind regulatory factors which protect the mRNA from degradation (17).

The 3′ UTR of the *ATM* gene, 3590 bp long, is among the longest known 3′ UTRs of mammalian mRNAs (31). It contains several AREs and its poly(A) tail is preceded by a canonical polyadenylation signal. Several additional potential polyadenylation signals are present in this region, including two canonical AATAAA. The use of these signals would generate various *ATM* mRNA species whose 3′ UTRs differ in length and in content of regulatory elements. Two shorter 3′ UTRs, identified by 3′ RACE, end with pairs of variants of polyadenylation signals with a lower expected polyadenylation activity: the two AGTAAA variants preceding the 1.6 kb product are expected to be ~30% efficient in polyadenylation activity compared to the canonical sequence, whereas the two GATAAA variants preceding the 1.1 kb product are expected to show only ~10% efficiency (32,33). The presence of two adjacent variant polyadenylation signals could have functional significance.

Since the actual size of the major *ATM* transcript observed on northern blots is ~13 kb (originally estimated by Savitsky *et al.* to be 12 kb; ref. 3), alternative transcripts containing 3′ UTRs which differ in size by several hundred nucleotides may not be resolved on such blots, and are most likely contained within the major *ATM* band. However, a particularly short 3′ UTR may account for the 10.5 kb minor transcript observed on northern blots from fibroblast RNA (3).

Transcripts of other genes have been shown to contain different 3′ UTRs which result from alternative polyadenylation site usage (34–38). Some of these examples suggest tissue-specific selection of a polyadenylation site and/or tissue-specific stability of the transcripts bearing different 3′ UTRs. The functional significance of the production of *ATM* transcripts that differ only in the length of their 3′ UTR is not clear. However, alternative polyadenylation could serve as a mechanism for regulating cellular levels and localization of different transcripts.

In conclusion, the extensive structural diversity of the 5′ and 3′ untranslated regions of *ATM* transcripts might reflect the need to carefully control the level of the ATM protein by promptly responding to physiological changes. Our data suggest that, as in other genes encoding regulatory proteins, this response is mediated by a multilevel regulatory process. Understanding the mechanisms involved in this complex regulation should lead to new insights into the biological role of the ATM protein. Experiments to elucidate the transcriptional and post-transcriptional regulatory mechanisms that control *ATM* expression are in progress.

ACKNOWLEDGEMENTS

This study was supported by research grants from the A-T Children's Project, the A-T Medical Research Foundation, the United States–Israel Binational Science Foundation and the National Institute of Neurological Disorders and Stroke (NS31763). This work was carried out in partial fulfillment of the requirements for the Ph.D. degree to K.S.

REFERENCES

1 Harnden, D.G. (1994) *Int. J. Radiat. Biol.* **66**, S13–S19.
2 Lavin, M.F. and Shiloh, Y. (1997) *Ann. Rev. Immunol.* **15**, 177–202.
3 Savitsky, K., Bar-Shira, A., Gilad, S., Rotman, G., Ziv, Y., Vanagaite, L., Tagle, D.A., Smith, S., Uziel, T., Sfez, S., *et al.* (1995) *Science* **268**, 1749–1753.
4 Savitsky, K., Sfez, S., Tagle, D.A., Ziv, Y., Sartiel, A., Collins, F.S., Shiloh, Y. and Rotman, G. (1995) *Hum. Mol. Gen.* **4**, 2025–2032.
5 Zakian, V.A. (1995) *Cell* **82**, 685–687.
6 Keith, C.T. and Schreiber, S.L. (1995) *Science* **270**, 50–51.
7 Jackson, S.P. (1995) *Curr. Biol.* **5,** 1210–1212.
8 Uziel, T., Savitsky, K., Platzer, M., Ziv, Y., Helbitz, T., Nehls, M., Boehm, T., Rosenthal, A., Shiloh, Y. and Rotman, G. (1996) *Genomics* **33,** 317–320.
9 Byrd, P.J., McConville, C.M., Cooper, P., Parkhill, J., Stankovic, T., McGuire, G.M., Thick, J.A. and Taylor, A.M.R. (1996) *Hum. Mol. Gen.* **5,** 145–149.
10 Byrd, P.J., Cooper, P.R., Stankovic, T., Kullar, H.S., Watts, G.D.J., Robinson, P.J. and Taylor, A.M.R. (1996) *Hum. Mol. Gen.* **5,** 1785–1791.
11 Imai, T., Yamauchi, M., Seki, N., Sugawara, T., Saito, T., Matsuda, Y., Ito, H., Nagase, T., Nomura, N. and Hori, T. (1996) *Genome Res.* **6**, 439–447.
12 Gilad, S., Khosravi, R., Shkedy, D., Uziel, T., Ziv, Y., Savitsky, K., Rotman, G., Smith, S., Chessa, L., Jorgensen, T.J., *et al.* (1996) *Hum. Mol. Gen.* **5**, 433–439.
13 Huang, X. (1994) *CABIOS* **10**, 227–235.
14 Meyuhas, O., Avni, D. and Shama, S. (1996) In Hershey, J.W.B., Mathews, M.B. and Sonenberg, N. (eds), *Translational Control.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 363–388.
15 Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125–8148.
16 Kozak, M. (1989) *J. Cell Biol.* **108**, 229–241.
17 Sachs, A.B. (1993) *Cell* **74**, 413–421.
18 Chen, C.Y.A. and Shyu, A.B. (1995) *Trends Biochem. Sci.* **20**, 465–470.
19 Hentze, M.W. (1995) *Curr. Opin. Cell Biol.* **7**, 393–398.
20 Morris, D.R. (1995) *Prog. Nucleic Acid Res. Mol. Biol.* **51**, 339–363.
21 Kozak, M. (1991) *J. Cell Biol.* **115**, 887–903.
22 Kozak, M. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2850–2854.
23 Geballe, A.P. (1996) In Hershey, J.W.B., Mathews, M.B. and Sonenberg, N. (eds), *Translational Control.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 173–197.
24 Kozak, M. (1992) *Annu. Rev. Cell Biol.* **8**, 197–225.
25 Jackson, R.J. (1996) In Hershey, J.W.B., Mathews, M.B. and Sonenberg, N. (eds), *Translational Control*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 71–112.
26 Bernstein, J., Sella, O., Le, S.Y. and Elroy-Stein, O. (1997) *J. Biol. Chem.* **272**, 9356–9362.
27 Ayoubi, T.A.Y. and Van de Ven, W.J.M. (1996) *FASEB J.* **10**, 453–460.
28 Jackson, R.J. (1993) *Cell* **74**, 9–14.
29 Decker, C.J. and Parker, R. (1995) *Curr. Opin. Cell Biol.* **7**, 386–392.
30 Ross, J. (1996) *Trends Genet.* **12**, 171–175.
31 Pesole, G., Fiormarino, G. and Saccone, C. (1994) *Gene* **140**, 219–225.
32 Sheets, M., Ogg, S.C. and Wickens, M.P. (1990) *Nucleic Acids Res.* **18**, 5799–5805.
33 Wickens, M. (1990) *Trends Biochem. Sci.* **15**, 277–281.
34 LeRoith, D. and Roberts, C.T. (1993) *Ann. NY Acad. Sci.* **692**, 1–9.
35 Code, R.J. and Olmsted, J.B. (1992) *Gene* **122**, 367–370.
36 Lin, B., Rommens, J.M., Graham, R.K., Kalchman, M., MacDonald, H., Nasir, J., Delaney, A., Goldberg, Y.P. and Hayden, M.R. (1993) *Hum. Mol. Gen.* **2**, 1541–1545.
37 Qian, J.F., Lazar-Wesley, E., Breugnot, C. and May, E. (1993) *Gene* **132**, 291–296.
38 Lu, Y. and Riegel, A.T. (1994) *Gene* **145**, 261–265.