The human ubiquitin-52 amino acid fusion protein gene shares several structural features with mammalian ribosomal protein genes

Rohan T.Baker+ and Philip G.Board*
Human Genetics Group, Division of Clinical Sciences, John Curtin School of Medical Research, Australian National University, PO Box 334, Canberra, ACT 2601, Australia

Received December 13, 1990; Accepted January 28, 1991 EMBL accession nos X56997, X56998, X56999

ABSTRACT

Complementary DNA clones encoding ubiquitin fused to a 52 amino acid tail protein were isolated from human placental and adrenal gland cDNA libraries. The deduced human 52 amino acid tail protein is very similar to the homologous protein from other species, including the conservation of the putative metal-binding, nucleic acid-binding domain observed in these proteins. Northern blot analysis with a tail-specific probe indicated that the previously identified UbA mRNA species most likely represents comigrating transcripts of the 52 amino acid tail (UbA_{52}) and 80 amino acid tail (UbA_{80}) ubiquitin fusion genes. The UbA_{52} gene was isolated from a human genomic library and consists of five exons distributed over 3400 base pairs. One intron is in the 5′ non-coding region, two interrupt the single ubiquitin coding unit, and the fourth intron is within the tail coding region. Several members of the *Alu* family of repetitive DNA are associated with the gene. The UbA_{52} promoter has several features in common with mammalian ribosomal protein genes, including its location in a CpG-rich island, initiation of transcription within a polypyrimidine tract, the lack of a consensus TATA motif, and the presence of Sp1 binding sites, observations that are consistent with the recent identification of the ubiquitin-free tail proteins as ribosomal proteins. Thus, in spite of its unusual feature of being translationally fused to ubiquitin, the 52 amino acid tail ribosomal protein is expressed from a structurally typical ribosomal protein gene.

INTRODUCTION

Ubiquitin is a small eukaryotic protein that exhibits extreme evolutionary conservation, with 71 of its 76 residues invariant over a broad range of species from yeast to man (reviewed in refs 1,2). The genes encoding ubiquitin exhibit two unique structural arrangements, which have also been strongly conserved (1). The polyubiquitin gene consists of tandem repeats of the 228 base pair (bp) ubiquitin coding unit in a head-to-tail spacerless array. The number of coding units varies considerably between species, and intraspecies variation is also observed in several organisms that have more than one polyubiquitin locus. The second structural type is the ubiquitin fusion gene, which encodes a single ubiquitin moiety fused to an unrelated 'tail' protein. Two sub-types can be identified by differences in the length and the sequence of the encoded tail. However, both tail proteins exhibit similarities such as a high proportion of basic residues, a putative nuclear localisation signal, and a cysteine-rich motif common to some nucleic acid binding proteins (3,4). Clues as to the function of these tail proteins have only recently been obtained: in the ubiquitin-free form, both are ribosomal proteins, with the small (52 residue) tail residing in the large subunit, while the large tail (76 or 80 residues) is a component of the small subunit (5,6). The fusion of ubiquitin to the N-terminus of these ribosomal proteins apparently increases the efficiency of their incorporation into the ribosome (5).

Human ubiquitin genes constitute a multigene family and are transcribed to produce mRNAs of approximately 600, 1000 and 2500 nucleotides (nt), termed UbA, UbB and UbC respectively (3,7). The UbC mRNA is transcribed from a nine coding unit polyubiquitin gene *UbC* (7), although unequal crossovers at this locus have resulted in *UbC* alleles containing only seven or eight coding units (8). The UbB subfamily is composed of: (i) a three coding unit polyubiquitin gene *UbB* containing a 715 bp intron within its 5′ non-coding region (9); (ii) at least three processed (i.e., intronless) pseudogenes (9,10); and (iii) a four coding unit non-processed (intron-containing) pseudogene (11). The UbA mRNA has been ascribed to the 80 amino acid tail ubiquitin fusion gene based on tail-specific hybridisation studies (3).

In this report we describe the nucleotide sequence and genomic organisation of a human gene, UbA_{52}, encoding the ubiquitin-52 residue tail fusion protein, and corresponding cDNA clones representing placental and adrenal gland transcripts. We also demonstrate specific hybridisation of a 52 residue tail-specific probe to the UbA mRNA species, suggesting that UbA represents

* To whom correspondence should be addressed

+ Present address: Department of Biology, Room 16-520, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

co-migrating transcripts encoding the ubiquitin-52- and ubiquitin-80-amino acid tail fusion proteins, which we propose be termed UbA_{52} and UbA_{80} respectively. The architecture of the UbA_{52} promoter strongly resembles that of mammalian ribosomal protein genes, consistent with the recent identification of the 52 amino acid tail as a component of the large ribosomal subunit (5).

MATERIALS AND METHODS

Recombinant libraries

A library consisting of partial *Sau*3AI-digested human genomic DNA cloned into the *Bam*HI sites of phage EMBL3A was the gift of Dr D. Anson. A human placental cDNA library in λgt11 and an adrenal gland cDNA library in λgt10 were from Clonetech. *Escherichia coli* host strains and manipulations involving recombinant phage were as described previously (9). Recombinant libraries were screened as described previously (12).

DNA manipulation

Routine procedures involving recombinant DNA were as described previously (13). Restriction maps of human genomic DNA inserts were determined as described elsewhere (13,14). Nucleotide sequences were determined by the chain termination method (15) employing the M13-mp phage derivatives (16).

Northern blot analysis

RNA was prepared (17) from lymphocytes purified by density gradient centrifugation through Lymphoprep (Nyegaard, Oslo) and from term placenta. Aliquots (10 μg) were glyoxylated and electrophoresed (13) prior to transfer to a nylon membrane (GeneScreen Plus, Du Pont) employing 10 mM NaOH as the transfer solution (18; K. Reed, pers. commun.). Membranes were hybridised according to the manufacturer's recommended protocols. Probes were generated by primer extension (19) of M13-mp phage subclones.

RESULTS

Isolation of a placental UbA_{52} cDNA clone

During the isolation of the human three coding unit polyubiquitin gene *UbB* (9), several clones were obtained that contained non-UbB ubiquitin sequences. Sequence analysis of one such clone (termed EHD5; unpublished data) revealed a pseudogene encoding ubiquitin plus a C-terminal extension homologous to the 52 amino acid 'tail' proteins of the yeast *UBI1* and *UBI2* genes (4). A probe derived from the tail-like coding region of EHD5 was used to screen a human placental cDNA library to enable characterisation of the human tail protein. Screening of 250,000 phage resulted in one repeatedly positive clone, PLUb5, that contained an insert of ~1300 bp, considerably longer than expected for a mRNA encoding a 128 residue protein. Sequence analysis revealed that the PLUb5 insert consisted of two cDNA clones fused head-to-head, with a poly(A) tail at each end (Fig. 1). One cDNA of 800 bp was found by computer-assisted analysis of the GenBank database to be a placental lactogen hormone (chorionic somatomammotropin) cDNA (20). The other cDNA (501 bp) had an open reading frame of 128 codons, coding for one 76 residue ubiquitin moiety followed by a 52 residue tail protein (Fig. 1) that was 81% identical to the yeast UBI1/UBI2 tail proteins (ref 4 and Fig. 2). Notably, the positions of the

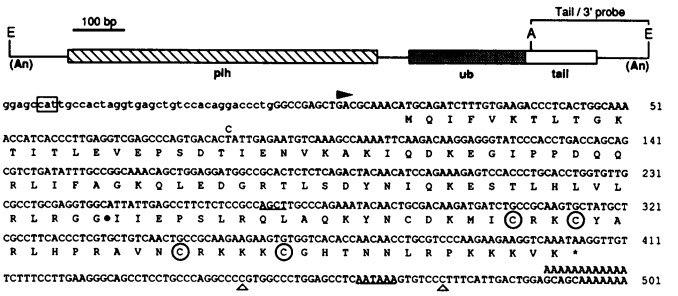

Figure 1. Structure and nucleotide sequence of UbA_{52} cDNA clones PLUb5 and ADUb2. Top: Structure of PLUb5 cDNA. Boxes represent coding sequences for placental lactogen hormone (plh), ubiquitin (ub) and tail (tail) proteins. Positions of poly (A) tails are shown by (An). The *Alu*I (A) and *Eco*RI (E) sites used to generate a tail-coding/3′ non-coding region specific probe are indicated. Bottom: Nucleotide sequence of the PLUb5 and ADUb2 cDNAs. UbA_{52} sequence is in upper case, plh sequence (partial) in lower case. The complement of the plh initiation codon is boxed. Numbering is from the first non-plh nucleotide. The ADUb2 sequence degins at position 10 (arrowhead) and is shown above PLUb5 only where it differs from it. The *Alu*I site used to construct the tail/3′ specific probe, and the polyadenylation signal are underlined. The translation is given underneath the sequence. The asterisk is the stop codon, and the junction between ubiquitin and tail proteins is shown by a filled circle. Cysteine residues comprising the zinc finger motif (see text) in the tail protein are circled. Open triangles point to bases absent from a partial leukocyte UbA_{52} cDNA sequence (24), most likely due to sequencing error (G. Salvesen, pers. commun.).

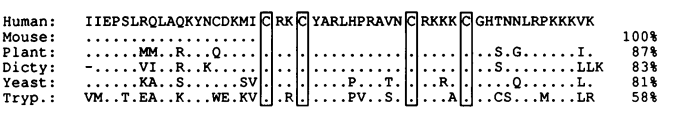

Figure 2. Comparison of tail protein sequences. Tail protein sequences from the organisms listed at left are compared to the human protein sequence (top line) given in the standard one-letter code. Only different residues are shown: identity to the human sequence is shown by a dot. A dash indicates a gap introduced into the *Dictyostelium discoideum* sequence ('Dicty') to maximise alignment. Percentage sequence identity to the human protein is given on the right. The invariant cysteine residues comprising the nucleic acid binding domain (21,22) are boxed. The first 19 residues of the mouse tail protein are deduced from a partial cDNA clone identified by St John et al (1986). Other sequences are from *A. thaliana* (Plant): ref 41; *D. discoideum* (Dicty): ref 44; *S. cerevisiae* (Yeast): ref 4; and *T. cruzi* (Tryp.): ref 39.

cysteine residues comprising the putative 'zinc finger' metal-binding, nucleic acid-binding domain (21,22) identified by Öazkaynak et al (4) in the yeast tail proteins are absolutely conserved in the PLUb5-encoded protein (Fig. 2).

The PLUb5 insert contains 18 bp between the lactogen hormone cDNA and the ubiquitin initiation codon, which presumably represent the 5′ non-coding region. The termination codon is followed by a 90 bp 3′ non-coding region and a 7 bp poly(A) tail, 27 bp downstream of the AATAAA polyadenylation signal (23). It was subsequently learned that this library was constructed from cDNAs of greater than 800 bp (Clonetech, pers. commun.). Thus a DNA complementary to the UbA_{52} mRNA of ~600 nt could only be present as a cloning artefact such as has occured in PLUb5, presumably arising during the ligation of *Eco*RI linkers during library construction. Therefore the observed low frequency of 1 in 250,000 clones may not be representative of the relative abundance of UbA_{52} mRNA in the placenta.

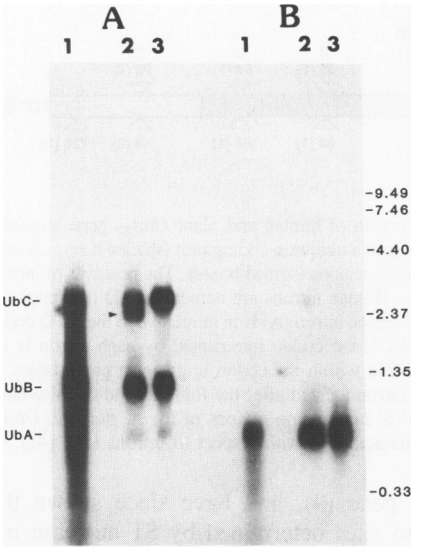

Figure 3. UbA$_{52}$ Northern blot analysis. Total RNA from human placenta (lane 1), freshly prepared lymphocytes (lane 2) and cultured lymphocytes (lane 3) was glyoxylated, electrophoresed, transferred to a nylon membrane and hybridised with a ubiquitin coding unit probe (A) or a *UbA$_{52}$* tail-coding/3′ non-coding region probe (B). Hybridising species are identified UbA, UbB and UbC (7). The placental sample is partially degraded but clearly exhibits the UbA$_{52}$ species. Size markers at the right are in thousands of nucleotides. The individual in lane 2 exhibits a length variation in the UbC transcript (arrowed) due to a polymorphism in the number of ubiquitin coding units per *UbC* allele, and is discussed elsewhere (8).

Isolation of adrenal gland *UbA$_{52}$* cDNA clones

A 240 bp *Alu*I/*Eco*RI tail-specific PLUb5 fragment containing sequences 3′ of the eighth tail codon (Fig. 1) was used to screen an adrenal gland cDNA library to isolate a *UbA$_{52}$* cDNA not originating from a cloning artefact. This probe hybridised to approximately 100 of 50,000 clones screened. Of 10 selected rescreened clones, three were chosen for sequence analysis. The clone with the longest insert, ADUb2, contained a 495 bp cDNA with 9 bp of 5′ non-coding region and differed from PLUb5 at two positions (Fig. 1). The first difference was a silent change in the 22nd codon of the ubiquitin coding region: threonine is encoded by ACT in PLUb5, and by ACC in ADUb2. This difference could reflect either allelic variation or an error arising during cDNA construction. The second difference was the site of polyadenylation: ADUb2 was polyadenylated 6 bp closer to the AATAAA signal than was PLUb5. The other two adrenal gland cDNAs sequenced were polyadenylated at the same position as ADUb2, but were partial cDNAs and were not informative on the difference seen at codon 22 (not shown). PLUb5 does not appear to represent an erroneously polyadenylated transcript, as a recently reported partial *UbA$_{52}$* cDNA from a leukemic cell line is also polyadenylated at this site (24). Alignment of this cDNA with PLUb5 reveals two sequence discrepancies in the 3′ non-coding region (Fig. 1), that are most likely due to sequencing errors in the leukocyte sequence (24; G. Salvesen, pers. commun.)

UbA$_{52}$ northern blot analysis

The identity of the *UbA$_{52}$* cDNAs as the products of a UbA subfamily gene was confirmed by the specific hybridisation of the PLUb5 tail-specific probe (see above) to the mRNA species previously identified as UbA (7). Northern blot analysis was conducted on RNA isolated from placenta, freshly prepared lymphocytes, and from a transformed lymphocyte cell line. Hybridisation with a ubiquitin coding region probe derived from a *UbB* polyubiquitin cDNA (9) revealed the UbA, UbB and UbC species (7; Fig. 3A). Hybridisation of a parallel northern blot with the tail-specific probe uniquely identified the UbA species (Fig. 3B). However, Lund et al (3) had previously observed specific hybridisation of the ubiquitin-80 amino acid tail fusion cDNA to the UbA mRNA. It thus appears that the UbA species represents two different co-migrating mRNAs. Hence it is proposed that the ubiquitin-52 amino acid tail species represented by PLUb5 and ADUb2 be termed *UbA$_{52}$*, and the 80 amino acid tail fusion species (3) would become *UbA$_{80}$*. This northern blot analysis also revealed a length polymorphism of the *UbC* mRNA (Fig. 3A, lane 2) which is described elsewhere (8).

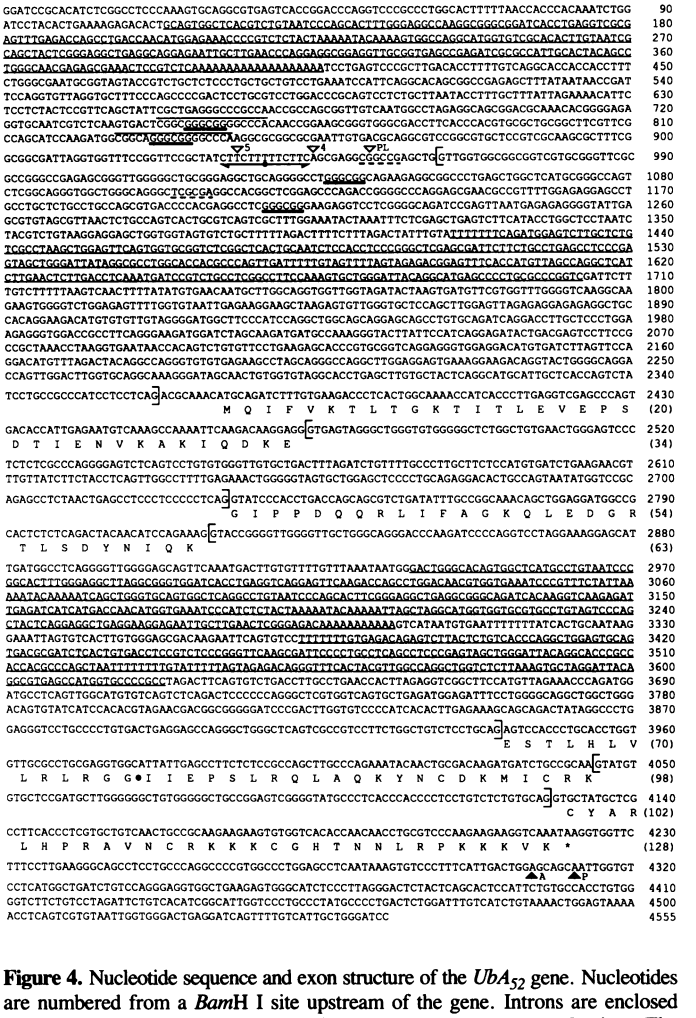

Figure 4. Nucleotide sequence and exon structure of the *UbA$_{52}$* gene. Nucleotides are numbered from a *Bam*H I site upstream of the gene. Introns are enclosed by square brackets positioned at splice donor ([) and acceptor (]) sites. The translation is shown below the sequence and amino acids are numbered in parentheses. The asterisk is the stop codon, and the junction between ubiquitin and tail proteins is shown by a filled circle. Sequences matching the core Sp1 promoter consensus (see text) are heavily underlined. The 17 bp direct repeat containing the two upstream Sp1 elements is overlined. Open triangles indicate transcription start sites observed in UbA$_{52}$ processed pseudogenes λUA4 ('4'), EHD5 ('5') (unpublished data), and the PLUb5 cDNA ('PL'). The 13 nt palindromic pyrimidine tract is underscored with a double headed arrow. Cleavage sites for restriction enzymes *Eag*I and *Nru*I are shown by a dashed underline. Filled triangles indicate polyadenylation sites employed in adrenal gland (A) and placental (P) cDNA clones. *Alu* repetitive DNA elements are underlined.

The human UbA_{52} gene

The PLUb5 UbA_{52} tail-specific probe was used to screen a human genomic library, resulting in the isolation of two clones containing different genomic inserts. One clone (termed λUA4) was found by sequence analysis to contain a UbA_{52} processed pseudogene (unpublished data). The second clone, λUA1, contained the UbA_{52} gene as described below.

Southern hybridisation analysis of λUA1 and its subclone pUA1.1 (not shown) indicated that the UbA_{52} cDNA-homologous region was distributed over more than 2 kb of DNA, suggesting the presence of introns within the gene. Determination of 4.55 kb of nucleotide sequence followed by comparison with the PLUb5 cDNA revealed firstly that λUA1 contains a *bona fide* gene, termed UbA_{52}, and secondly that the transcribed region is distributed over 5 exons (Fig. 4). Exon 1 contains 29 bp of 5′ non-coding region (see below). Exon 2 is 111 bp long, containing 8 bp of 5′ non-coding region and ubiquitin codons 1 through 34.33. Exon 3 (87 bp) contains ubiquitin codons 34.33 to 63.33, and exon 4 (103 bp) contains ubiquitin codons 63.33 to 76 and tail codons 1 to 21.67. Exon 5 contains tail codons 21.67 to 52 and the 3′ non-coding region. Introns A through D are respectively 1400, 259, 1122 and 84 bp in length. All splice junctions confer with the 'GT-AG' rule and match the consensus sequences for these sites (25).

The protein encoded by UbA_{52} is identical to the cDNA-encoded proteins. The gene matches the adrenal gland cDNA sequence at ubiquitin codon 22 (ACC) rather than the placental ACT (Fig. 4, nt 2434). However, the gene differs from both cDNAs 4 bp downstream of the termination codon, containing a G (gene, nt 4226) instead of a T (cDNAs). As discussed above, this difference may reflect allelic variation or a cDNA synthesis/cloning error. In addition to the canonical AATAAA polyadenylation signal (23) the polyadenylation region contains sequences matching the consensus elements CAYTG (CATTG, nt 4296) and the T/G cluster (TGGTGTCCT, nt 4315), which have been implicated in mRNA 3′ end formation (26,27).

Four members of the *Alu* family of repetitive DNA are associated with UbA_{52} (Fig. 4). A complete *Alu* repeat is present 528 bp upstream of exon 1 and another within intron A, respectively 89 and 87% identical to the consensus (28). Intron C contains two members, one of which (88% identical) has suffered a 38 bp deletion in the first monomer, while the other is a complex repeat, consisting of one *Alu* first monomer (89%) followed immediately by a complete *Alu* unit (84%) which has suffered a 68 bp deletion in its second monomer. This whole complex member is flanked by a short direct repeat. These two intron C *Alu* repeats comprise more than half the intron, are in opposite orientations separated by 77 bp, and thus form a large inverted repeat within the intron.

UbA_{52} promoter and 5′ non-coding region

The placental UbA_{52} cDNA clone PLUb5 contains 18 bp of 5′ non-coding region, of which the first 8 bp immediately upstream of the initiation codon are included in exon 2 of UbA_{52} (Fig. 4). The other 10 bp of cDNA 5′ non-coding region are present 1400 bp upstream in exon 1. Comparison of the gene sequence with the two UbA_{52} pseudogenes around this region (unpublished data) suggests that exon 1 is at least 29 bp in length: homology to the two processed pseudogenes ceases 18 bp and 29 bp upstream of the splice donor site (Fig. 4). We have previously used limits of homology between gene and processed pseudogene sequences to identify the transcription initiation site of the *UbB*

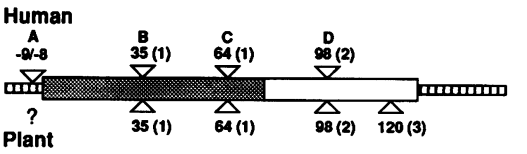

Figure 5. Comparison of human and plant UbA_{52} gene structure. The UbA_{52} mRNA is divided into a ubiquitin coding unit (shaded box), tail coding unit (open box), and non-coding regions (striped boxes). The positions of introns are indicated by open triangles. Human introns are named A to D (see text). The position of the 5′ non-coding region intron (A) is in nt relative to the ATG codon. For coding-region introns (B-D), the codon interrupted by each intron is numbered. The position of the intron within the codon is given in parentheses, whereby 1, 2, and 3 signify an intron located after the first, second or third base of the codon respectively. The 5′ non-coding regions of the *A. thaliana* UbA_{52} homologues have not been characterised with respect to introns ('?') (41).

polyubiquitin gene (9), and have since shown that this site corresponds to sites determined by S1 nuclease mapping and sequencing of full-length cDNA clones (29). Thus it is probable that transcription initiates at or around nucleotide 934 (Fig. 4). In this respect, Finley et al (5) have noted that the yeast ubiquitin fusion genes exhibit several sequence features common to yeast ribosomal protein (rp) genes, such as intron positioning and the presence of rp gene-specific promoter elements. We thus looked for features of UbA_{52} that were common to known mammalian rp genes. The most consistent features of the latter are: (i) a small first exon and 5′ non-coding region; (ii) lack of a consensus TATA promoter; (iii) initiation of transcription within a pyrimidine tract (often palindromic) embedded in a CpG-rich island; and (iv) the presence of Sp1 binding sites (30,31). UbA_{52} conforms to all of these features. Exon 1 is 29 bp in length and the untranslated leader is 37 bp. The closest consensus TATA sequence is positioned 406 bp upstream (nt 528, Fig. 4), too distant from any known transcribed sequence to be of significance. However, the transcription start site at position 934 lies within a palindromic 13 bp pyrimidine tract (Fig. 4). Inspection of the sequence around this site suggests a CpG-rich island: the 585 bp region from −304 to +281 relative to position 934 (nt 630 to 1214, Fig. 4) has a G + C content of 70%, no underrepresentation of the CpG dinucleotide, and also contains cleavage sites for the 'CpG-rich island specific' restriction enzymes *Eag*I (nt 952) and *Nru*I (nt 1105). These three features are considered indicative of CpG-rich islands (31,32). This region also contains four elements matching the consensus Sp1 binding site (33), two upstream and two downstream of exon 1 (Fig. 4). The two upstream sequences are 9-of-10 matches to the expanded Sp1 consensus (33) and are part of a larger direct repeat as follows:

nt 743	CGGC-GGGCGGGGCCCA
nt 827	CGGCAGGGCGGGGCCCA

In addition, the 19 bp surrounding the Sp1 box at nucleotide 747 matches a 20 bp Sp1-containing element in the first intron of the human α1(I) collagen gene (34) as follows:

UbA_{52}, nt 739	GACTCGGC-GGGCGGGGcCC
Collagen	GACTCGGCGGGGCGGGGtCC

The location of these Sp1 binding sites in close proximity to known UbA_{52} transcribed regions suggests that transcription of this gene may be Sp1 regulated.

DISCUSSION

UbA represents two mRNAs

The UbA_{52} tail-specific probe uniquely identifies the UbA mRNA species on a Northern blot (Fig. 2). As the 80 amino acid tail cDNA also specifically hybridises to the same species (3), it thus appears that UbA represents co-migrating UbA_{52} and UbA_{80} transcripts. However, our studies have been limited to placental and lymphocyte tissues, whereas Lund et al (3) analysed liver and a mammary carcinoma cell line. As none of these tissues coincide, an alternate possibility of tissue-specific expression of UbA_{52} and UbA_{80} cannot be excluded at this stage. However, this must be considered extremely unlikely, as these transcripts encode ribosomal proteins (5,6), and would thus be expected to be co-ordinately expressed in all actively translating tissues. Although the UbA_{80} mRNA encodes an additional 28 amino acids, it has a very short (28 nt) 3′ non-coding region (3). Thus, excluding the 5′ non-coding regions (absent from the reported UbA_{80} cDNA), the unpolyadenylated lengths of the UbA_{52} and UbA_{80} mRNAs would be 477 and 499 nt respectively. Assuming similar 5′ non-coding region and poly (A) tail lengths, a difference of 22 nt between two ~600 nt mRNAs would not be resolvable with the ubiquitin coding region probe (Fig. 2). Northern analysis employing both tail specific probes on the same tissues is required to firmly demonstrate the co-migration of the two transcripts.

UbA_{52} gene structure

UbA_{52} consists of 5 exons separated by 4 introns. The 128 codons specifying the ubiquitin-52 amino acid tail fusion protein are distributed approximately equally over exons 2 to 5, with exon 2 also containing 8 bp of 5′ non-coding region, and exon 5 containing the 3′ non-coding region of 84 or 90 bp. The length of the 3′ non-coding region varies with the tissue source of the cDNA: while the same polyadenylation signal is employed, the polyadenylation site in the placental cDNA PLUb5 and a leukocyte UbA_{52} cDNA (24) is 6 bp downstream from that used in the three sequenced adrenal gland cDNAs. However, these five cDNA clones representing three tissues are too few in number to confirm either alternate polyadenylation sites in all tissues, or tissue-specific polyadenylation sites.

UbA_{52} shares several structural features with mammalian ribosomal protein (rp) genes at its 5′ end, including its location within a CpG-rich island, a short first exon and 5′ non-coding region, lack of a consensus TATA promoter, the presence of Sp1 binding sites, and the (putative) initiation of transcription within a palindromic pyrimidine stretch. Given that UbA_{52} encodes a ribosomal protein fused to the C-terminus of ubiquitin, these features are not totally unexpected. Indeed, Finley et al (5) have observed that the yeast ubiquitin fusion genes share structural features with yeast rp genes. UbA_{52} is also similar to mammalian rp genes in that it is a member of a multigene family whose other members are processed pseudogenes. Southern hybridisation and nucleotide sequence analysis suggests a large UbA_{52} subfamily containing approximately eight processed pseudogenes and only one expressed, intron-containing gene, UbA_{52} (R.T.B. and P.G.B., manuscript in preparation).

The Sp1 transcription factor is involved in the transcription of the 'housekeeping' genes; i.e., those which are constitutively expressed in a wide variety of tissues (33). The ribosomal protein genes are a good example of housekeeping genes, and Sp1 binding sites are found upstream of some, but not all, mammalian rp genes (31). It has recently been shown that an Sp1 box positioned 161 bp upstream of the mouse rpS16 gene binds the Sp1 transcription factor, resulting in a 2.5 fold increase in rpS16 transcription (35). By analogy to this latter finding, the two UbA_{52} Sp1 boxes spaced 93 and 178 bp upstream of exon 1 are located sufficiently close to UbA_{52} to influence its transcription. In addition, these Sp1 boxes are part of a larger repeat unit that is very similar to a functional Sp1 box in the human α1(I) collagen gene (34), two observations that reinforce their potential for involvement in UbA_{52} expression.

An interesting feature is the presence of four introns in UbA_{52}; one within the 5′ non-coding region and three interrupting the coding region. Several polyubiquitin genes contain a 5′ non-coding region intron positioned 5 to 11 bp upstream of the initiation codon (9,36,37). Conversely, introns are absent from the coding regions of all known polyubiquitin genes (4,7,9,36–39) except for the *Caenorhabditis elegans* 11 coding unit polyubiquitin gene, unusual in that the 1st, 4th, 7th and 10th coding units contain an identically-positioned ~50 bp intron (40). Notably, neither of the two introns within the UbA_{52} ubiquitin coding unit (codons 35 and 64) correspond in position to the *C. elegans* introns (codon 47). Ubiquitin fusion gene structure is less well characterised and has only been described for the yeast *S. cerevisiae* (4), the plant *Arabidopsis thaliana* (41), and for the 52 amino acid fusion gene from the parasitic protozoan *Trypanosoma cruzi* (39). The yeast *UBI3* gene and plant *UBQ5* and *UBQ6* genes (UbA_{80} homologues) and *T. cruzi FUS1* gene (UbA_{52} homologue) are intronless (except for the *trans*-spliced mini exon in *T. cruzi*), whereas the yeast *UBI1*/*UBI2* genes (UbA_{52} homologues) contain a single intron interrupting the third ubiquitin codon. This intron position is not conserved between the yeast and human ubiquitin-52 amino acid fusion genes. The recent determination of the structure of the *A. thaliana* UbA_{52} homologues, *UBQ1* and *UBQ2* (41), provides an interesting comparison. Although the plant 5′ non-coding region has not been characterised with respect to introns, the positions of the three UbA_{52} coding region introns are identical to corresponding introns in both copies of the plant gene. In addition, both plant genes contain one extra intron interrupting the tail coding region (Fig. 5). Thus the intron arrangement of the ubiquitin-52 amino acid tail fusion gene has been well conserved during the evolution of these higher eukaryotes.

A further feature of UbA_{52} exon organisation with respect to the unusual structural arrangement of the ubiquitin fusion genes is that the ubiquitin and tail proteins are not encoded by completely separate (sets of) exons, but that exon 4 encodes the 13 C-terminal residues of the former plus the first 22 residues of the latter. At face value it thus appears that the UbA_{52} fusion gene was not created by exon shuffling of two preformed independent genes, unless the loss of a putative intron separating the ubiquitin and tail portions of exon 4 is invoked. As the exon 4 homologue is structured identically in *A. thaliana* (41; Fig. 5), such an intron loss must predate the plant/animal divergence (ignoring the lower probability event of separate, identical intron losses since divergence). However, the plant gene coding region contains one extra intron compared to UbA_{52} (Fig. 5), and thus intron loss/generation during the evolution of this fusion gene in these species is not without precedent.

Features and functions of the tail protein

All known transcriptionally active ubiquitin genes encode fusion proteins, either of ubiquitin to itself to produce polyubiquitin, or to a non-ubiquitin tail protein, as with UbA_{52}. Thus ubiquitin

is always generated by post-translational proteolytic processing. Initial observations of the highly basic amino acid composition and a nuclear translocation-type signal within the tail proteins suggested a nuclear location and perhaps a function as a carrier to transport ubiquitin to the nucleus for its known conjugation to histones (3). However, the identification of the metal-binding, nucleic acid-binding 'zinc finger' domain within both tail types (4) and the high level of evolutionary conservation of the tail proteins (Fig. 2) are indicative of a specific function(s). Recently, the de-ubiquitinated tail proteins have been identified as ribosomal proteins, with the small and large tails present in the large and small subunits, respectively (5,6). Presumably, the cysteine-rich zinc finger domain identified in both tail proteins is involved in RNA binding. The synthesis of these, but not any other, ribosomal proteins as C-terminal fusions to ubiquitin in apparently every eukaryotic organism raises questions as to the function of this structural arrangement. Finley et al (5) have demonstrated that, at least for the yeast *UBI3* gene (*UbA_{80}* homologue), the presence of ubiquitin at the N-terminus of this ribosomal protein greatly facilitates its incorporation into the ribosome. However, the fusion protein is known to be very rapidly processed in yeast *in vivo* to ubiquitin and free tail (ribosomal) protein, perhaps even co-translationally, and the intact fusion protein generally cannot be detected (5,42). Thus the facilitative function of the ubiquitin fusion must also be exerted co-translationally, or be due to a small (undetectable) fraction of the fusion protein that remains unprocessed. In either case, the structure-function relationship of the ubiquitin fusion genes presents a very interesting evolutionary question.

ACKNOWLEDGEMENTS

We thank Bonnie Bartel for critically reviewing the manuscript. R.T.B. was supported by a Commonwealth Postgraduate Research Award.

REFERENCES

1. Schlesinger,M.J. and Bond, U. (1987) *Oxf. Surv. Euk. Genes*, **4**, 77–91.
2. Sharp,P.M. and Li,W.-H. (1987) *Trends Ecol. Evol.*, **2**, 328–332.
3. Lund,P.K., Moats-Staats,B.M., Simmons,J.G., Hoyt,E., D'Ercole,A.J., Martin,F. and Van Wyk,J.J. (1985) *J. Biol. Chem.*, **260**, 7609–7613.
4. Özkaynak,E., Finley,D., Solomon,M.J. and Varshavsky,A. (1987) *EMBO J.*, **6**, 1429–1439.
5. Finley,D., Bartel,B., and Varshavsky,A. (1989) *Nature*, **338**, 394–401.
6. Redman,K. and Rechsteiner,M. (1989) *Nature*, **338**, 438–440.
7. Wiborg,O., Pedersen,M.S., Wind,A., Berglund,L.E., Marcker,K.A. and Vuust,J. (1985) *EMBO J.*, **4**, 755–759.
8. Baker,R.T. and Board,P.G. (1989) *Am. J. Hum. Genet.*, **44**, 534–542.
9. Baker,R.T. and Board,P.G. (1987a) *Nucl. Acids Res.*, **15**, 443–463.
10. Baker,R.T. and Board,P.G. (1987b) *Nucl. Acids Res.*, **15**, 4352.
11. Cowland,J.B., Wiborg,O. and Vuust,J. (1988) *FEBS Lett.*, **231**, 187–191.
12. Benton,W.D. and Davis,R.W. (1977) *Science*, **196**, 180–181.
13. Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) Molecular Cloning: A Laboratory Manual. Cold Spring Harbour Laboratory, Cold Spring Harbor.
14. Baker,R.T. and Board,P.G. (1988) *Nucl. Acids Res.*, 16, 1198.
15. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
16. Messing,J. (1983) *Methods Enzymol.*, 101, 20–79.
17. Chomczynski,P. and Sacchi,N. (1987) *Anal. Biochem.*, **162**, 156–159.
18. Reed,K.C. and Mann,D.A. (1985) *Nucl. Acids Res.*, **13**, 7207–7221.
19. Burke,J.F. 1984. *Gene*, **30**, 63–68.
20. Seeburg,P.H. (1982) *DNA*, **1**, 239–249.
21. Miller,J., McLachlan,A.D. and Klug,A. (1985) *EMBO J.*, **4**, 1609–1614.
22. Berg,J.M. (1986) *Science*, **232**, 485–487.
23. Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **263**, 211–214.
24. Salvesen,G., Lloyd,C. and Farley,D. (1987) *Nucl. Acids Res.*, **15**, 5485.
25. Breathnach,R. and Chambon,P. (1981) *Ann. Rev. Biochem.*, **50**, 349–383.
26. Berget,S.M. (1984) *Nature*, **309**, 179–182.
27. McLauchlan,J., Gaffney,D., Whitton,J.L. and Clements,J.B. (1985) *Nucl. Acids Res.*, **13**, 1347–1368.
28. Kariya,Y., Kato,K., Hayashizaki,Y., Himeno,S., Tarui,S. and Matsubara,K. (1987) *Gene*, **53**, 1–10.
29. Baker,R.T. (1988) PhD thesis, Australian National University, Canberra, Australia.
30. Mager,W.H. (1988) *Biochim. Biophys. Acta*, **949**, 1–15.
31. Huxley,C. and Fried,M. (1990) *Nucl. Acids Res.*, 18, 5353–5357.
32. Lindsay,S. and Bird,A.P. (1987) *Nature*, **327**, 336–338.
33. Kadonaga,J.T., Jones,K.A. and Tijan,R. (1986) *Trends Biochem. Sci.*, **11**, 20–23.
34. Bornstein,P., McKay,J., Morishima,J.K., Devarayalu,S. and Gelinas,R.E. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 8869–8873.
35. Hariharan,N., and Perry,R.P. (1989) *Nucl. Acids Res.*, 17, 5323–5337.
36. Bond,U. and Schlesinger,M.J. (1986) *Mol. Cell. Biol.*, 6, 4602–4610.
37. Lee,H., Simon,J.A. and Lis,J.T. (1988) *Mol. Cell. Biol.*, **8**, 4727–4735.
38. Giorda,R. and Ennis,H.L. (1987) *Mol. Cell. Biol.*, **7**, 2097–2103.
39. Swindle,J., Ajioka,J., Eisen,H., Sanwal,B., Jacquemot,C., Browder,Z. and Buck,G. (1988) *EMBO J.*, **7**, 1121–1127.
40. Graham,R.W., Jones,D. and Candido,E.P.M. (1989) *Mol. Cell. Biol.*, **9**, 268–277.
41. Callis,J., Raasch,J.A. and Vierstra,R.D. (1990) *J. Biol. Chem.*, **265**, 12486–12493.
42. Monia,B.P., Ecker,D.J., Jonnalagadda,S., Marsh,J., Gotlib,L., Butt,T., and Crooke,S.T. (1989) *J. Biol. Chem.*, **264**, 4093–4103.
43. St. John,T., Gallatin,W.M., Siegelman,M., Smith,H.T., Fried,V.A., and Weissman,I.L. (1986) *Science*, **231**, 845–850.
44. Müller-Taubenberger,A., Westphal,M., Jaeger,E., Noegel,A. and Gerisch,G. (1988) *FEBS Lett.*, **229**, 273–278.