Human link protein gene: structure and transcription pattern in chondrocytes

Jayesh DUDHIA,* Michael T. BAYLISS and Timothy E. HARDINGHAM

Kennedy Institute of Rheumatology, 6 Bute Gardens, Hammersmith, London W6 7DW, U.K.

We have examined the genomic organization and the transcription unit for the human link protein gene from genomic clones and RNA prepared from human cartilage over a wide age range. Five exons cover the gene which is greater than 60 kbp. Primer extension and S1 nuclease protection analysis revealed transcription initiation to be 315 bases upstream from the translation initiation codon in RNA derived from cartilage samples ranging from fetal to 53 years of age. The first exon size therefore is 289 bp and examination of the 5' flanking sequence indicated a lack of a TATA box in close proximity to the

transcription start, although a TATAA-like motif (TCTAA) was present at –75 bp. Such a sequence at a similar distance can serve as a promoter in the chicken link protein gene. The large first exon of 289 bp is similar to that of the chicken but contrasts with that described previously for human (96 bp) and rat (62 bp). We also analysed human link protein mRNA by PCR for the presence of an alternatively spliced exon that is present in rat mRNA in low abundance, but could not detect such transcripts. Equine and porcine mRNA contained this spliced form but the results suggested that this was expressed as a rare transcript.

INTRODUCTION

Cartilage is a specialized connective tissue with a highly organized extracellular matrix. This matrix contains a dense fibrillar collagen network and abundant proteoglycan aggregates formed by the chondroitin sulphate/keratan sulphate proteoglycan, aggrecan, and hyaluronan [1,2]. Together these provide the cartilage with resilience to compression and good load-bearing properties. Link protein is a globular extracellular glycoprotein (molecular mass 40–48 kDa) which stabilizes the association between aggrecan and hyaluronan [3,4] by binding to both molecules via specific non-covalent interactions. Link protein contains three structural domains: an Ig-fold which interacts with the aggrecan G1 domain, and two proteoglycan tandem repeats (PTR) which bind to hyaluronan [5,6].

The aggrecan G1 domain contains Ig-fold and tandem-repeat motifs [7] which are similar in sequence and structure to those present in link protein [8]. Other members of the hyaluronan-binding proteoglycan family, versican and neurocan, also contain structurally related G1 domains [9] with which link protein may also interact. The presence of link protein in other non-cartilaginous tissues may thus be associated with the expression of these proteoglycans. Link protein is co-expressed with aggrecan and collagen type II in chondrocytes during chondrogenesis [10] but it is also present in aorta, sclera and lens capsule [11] and in the synovium [12].

From the cloned cDNA sequence [13,14] the deduced protein contains 354 amino acids encoded by a single gene. In cartilage matrix the protein exists predominantly as three forms, two (LP-1 and LP-2) that vary only in their glycosylation and a third form that is a proteolytic product (LP-3) that lacks a 13-amino-acid N-terminal peptide [15–18]. The released peptide is of interest as it is specifically taken up by several cell types and can stimulate proteoglycan synthesis in chondrocytes [19]. LP-3, although lacking this peptide, retains its properties of stabilizing aggrecan binding to hyaluronan. The relative abundance of LP-1, -2, and

-3 varies among different species and cartilage sources; older cartilage may also contain link protein species with further proteolytic damage [20,21]. No evidence for the expression as protein of alternate spliced forms of link protein mRNA has been reported.

The mRNA of link protein exists in transcripts of widely different size (1.2–5.1 kb) in all species examined, which involves particularly large variations in the length of the 3' untranslated region. In the chicken multiple polyadenylation sequences have been identified [22] that may account for this variation in size. Several transcription initiation start sites were also detected and the 5' untranslated region also varied as a result of alternative splicing [23]. In contrast, an investigation of the rat link protein gene found evidence of only a single transcription start site and some limited evidence from the human link protein gene suggested that it was similar to that of the rat [24]. In all species investigated the mRNA sequence coding for the translated protein was invariant, with the exception of a report [25] of a low-abundance alternatively spliced exon in rat mRNA.

In this study we have carried out a more detailed study of human link protein gene structure as our preliminary evidence of the transcription start site differed from that published. We have also localized the exon–intron boundaries from cosmid clones covering regions of the human gene not previously characterized.

MATERIALS AND METHODS

RNA preparation, primer extension and S1 nuclease protection

Total RNA was prepared from chondrocytes isolated from normal human femoral condyles from subjects aged 10, 19, 20, 26, 27 and 53 years, and from fetal femoral and tibial cartilage of 26 weeks gestation. Chondrocytes were released by collagenase digestion of the matrix [26], and freed cells plated on plastic dishes at a density of 1×10^7 cells/90-mm-diam. dish. Cells were cultured for 3 days in Ham's F12 medium containing 10% (v/v) fetal-calf serum. RNA was extracted using guanidine thiocyanate

Abbreviation used: PTR, proteoglycan tandem repeat.

* To whom correspondence should be addressed.

The nucleotide sequences described here have been submitted to EMBL/GenBank data library under the Accession numbers X78075 and X78076.

and acidic phenol [27]. RNA was also prepared from CaCo-2 cells (ATCC HTB37), a human colon adenocarcinoma cell line [28] that we confirmed by Northern blot to express link protein mRNA.

For primer extension, 10 µg of total RNA was denatured at 65 °C and immediately annealed to 1 nmol of ³²P-labelled primer in 100 mM Tris/HCl, pH 7.5, 1 mM EDTA, 400 mM NaCl at 30 °C for 2 h. Two primers whose sequences were reverse and complementary to the cDNA sequence [14] were used: 90/H (5'-GTCTCTTTGTGGCGCTGGATGG) and 91/B (5'-GGAAGCTGGAGATCAAGTGAAGCTAGCGCTCTTTC). The RNA was precipitated in ethanol and resuspended in 100 mM Tris/HCl, pH 8.3, 80 mM KCl, 12 mM MgCl₂, 10 mM dithiothreitol, 0.2 mM dNTP, and 200 units of M-MLV reverse transcriptase (BRL) and 10 units of RNAguard (Pharmacia) for elongation at 37 °C for 1 h. For S1 nuclease protection a forward and a reverse primer (90/H and 91/C respectively, see Figure 1) spanning the putative transcription start point identified by primer extension were used to generate a 185 bp probe by PCR amplification of a cosmid clone encoding the human link protein gene (see below). The DNA fragment was end-labelled with [γ-³²P]ATP using T4 polynucleotide kinase. Approx. 1 × 10⁶ c.p.m. of denatured probe was annealed with 10 µg of total RNA, samples were diluted 10-fold with S1 nuclease buffer (30 mM sodium acetate, pH 4.6, 100 mM NaCl, 1 mM ZnSO₄) and digested at 30 °C with 300 units of S1 nuclease (Pharmacia) for 1 h.

The reaction products of primer extension and S1 nuclease assays were extracted in phenol/chloroform (1:1, v/v) and precipitated in ethanol. Pellets were redissolved in 80 % formamide and aliquots (6 µl) electrophoresed on 8 M urea/6 % polyacrylamide sequencing gels.

Genomic cloning, DNA sequencing and PCR amplification

Link protein genomic clones were isolated from a human cosmid library (gift of Dr. T. H. Rabbitts, MRC Laboratory of Molecular Biology, Cambridge, U.K.) constructed in the Loris 6 vector using a partial *Bgl*II digest of human dermal skin fibroblast DNA. The library was screened using human link protein cDNA from the clone λ8.1D3 [14]. Dideoxynucleotide DNA sequencing was carried out on double-stranded DNA using Sequenase T7 DNA polymerase (US Biochemical) and [³²P]dATP. Sequence was determined in both directions using specific oligonucleotide

Figure 1 Nucleotide sequence at the 5'-end of the human link protein gene and first exon

Numbering of nucleotides starts at the transcription initiation site marked +1. The first exon is marked within the box. Relevant positions of primers used for primer extension and S1 protection analysis are underlined. A TATAA motif-like sequence has a double underline.

primers. PCR amplification was performed in 50 µl volumes using *Taq* DNA polymerase (Perkin-Elmer) in buffer conditions listed by the supplier, and 1 nmol of each primer in the Omnigene thermal cyclor (Hybaid Ltd).

RESULTS AND DISCUSSION

Genomic structure

To investigate the genomic structure of the human link protein gene a cosmid library of human DNA was screened with our full-length link protein cDNA [14]. Five clones were isolated and these were probed with a 100 bp DNA fragment generated by PCR amplification of the 5'-end of the link protein cDNA. A single 36 kbp clone, cos2.3B, hybridized positively with this probe. To analyse the cos2.3B clone further it was cleaved with *Pst*I and the fragments were subcloned into pBluescript KS and a subclone with a 5.1 kbp insert was identified with the 100 bp 5' probe by blot hybridization. Partial sequencing (Figure 1) of this subclone showed that it corresponded to the link protein cDNA sequence and part of the 5' flanking genomic region. This clone

Table 1 Summary of the exon-intron structure of the human link protein gene

Nucleotide sequences at the exon (uppercase letters) and intron (lowercase letters) junctions are shown. Numbers above the exon sequences are denoted from the first transcribed base in Figure 1 and continue in sequence in the link protein mRNA described in [14]. The consensus sequence is from [31,32].

Exon	Size (bp)	5' splice site	Intron	3' splice site
1	289	289		290
		ACAAGgtaagaaat..		..cctccatagTGAAG
2	126	415		416
		CCAAGgtaagcaat..		..taattatagCAGAA
3	372	787		788
		ACAAGgtgagat..		..tttgacagGTGTG
4	313	1100		1101
		CAATGgtaacgat..		..tctctacagGCGGT
5	> 3000			
Consensus sequence	C	a		cccccc c
		AGgt agt		n agG
	A	g		tttttt t

Figure 2 PCR amplification of cartilage RNA using primers spanning exons 2 and 3 of link protein

A sample (25 ng) of total RNA was primed with 1 ng of random hexamers to prepare cDNA using M-MLV reverse transcriptase, which was omitted in a control reaction. The two primers are shown in the schematic where primer 89/A lies within exon 2 and primer 89/C within exon 3. The arrows indicate the DNA fragment length amplified from cDNA, and does not include intron 2. The relative position of the rat exon 2 is denoted as RE2. Amplification was carried out under conditions (92 °C, 1 min; 52 °C, 1 min; 72 °C, 40 s; 35 cycles) that routinely could amplify 550 bp of cDNA. Aliquots were analysed on a 3% NuSieve agarose gel; size markers are *Pst*I digest of λ DNA. Lane 1, human RT-PCR; lane 2, human control; lane 3, porcine RT-PCR; lane 4, equine RT-PCR.

corresponded to the 5'-end of the parent cos2.3B clone and, as it contained additional sequence upstream from the cDNA untranslated region, it was likely to contain promoter elements of the link protein gene. The other clones, which failed to hybridize to this probe, therefore contained 3'-coding sequences and one of these, cos1.3A, was used for further 3' analysis studies.

The donor and corresponding acceptor splice sites (Table 1) within clones cos2.3B and cos1.3A were determined by double-stranded sequencing using specific oligonucleotide primers synthesized on the basis of the sequence known from the cDNA [14]. This analysis confirmed that the human link protein gene contained five exons, with each structural domain of the protein encoded within separate exons. Thus exon 1 contains most of the 5' untranslated region; exon 2 the translated protein leader sequence; exon 3 the Ig-fold domain; and exons 4 and 5 the B and B' loops of the PTR respectively. Exon 5 also encodes all of the 3' untranslated region which varies from 0.3 to 3.6 kbp. Southern-blot analysis (results not shown) suggests that the size of the gene is greater than 60 kbp, largely contributed by intron 1 which is greater than 25 kbp. The 5' untranslated region (315 bp) and 3' untranslated region (up to 3.6 kb) are unusually large, a feature also present in porcine [8] and chicken [22] link protein mRNA.

An unusual alternatively spliced mRNA has been previously reported in low abundance (% of total link protein mRNA) from rat chondrosarcoma [25], which predicted an extra translated protein sequence, although no evidence for the translated product was obtained. We tested for the presence of this extra exon by reverse-transcriptase-mediated PCR of human chondrocyte RNA using oligonucleotide primers specific to the link protein sequence and straddling the putative exon position. However, a comparable mRNA with an extra exon, shown in Figure 2, could not be detected in human RNA. Analysis of porcine and equine chondrocyte RNA (Figure 2, lanes 3 and 4) showed the presence of an extra band of the expected size, but this was only as a very minor component compared with the band generated by link protein transcripts that lacked this exon. Increasing the loading of PCR-amplified DNA of the human sample to the same intensity as the equine and porcine 369 bp

Figure 3 Northern-blot analysis of cartilage RNA

A sample (1 μ g) of total RNA per lane from adult cartilage was electrophoresed through 1.2% agarose gels with 2.2 M formaldehyde, and transferred to nylon membrane (Hybond N, Amersham). Antisense riboprobes labelled with [32 P]dCTP were prepared from cDNA fragments cloned into pBluescript and used as hybridization probes. Post-hybridization washes of membranes were under stringent conditions, 0.1 $\times$ SSC, 65 °C, before autoradiography at -70 °C for 2-4 h. Lane 1, porcine RNA; lane 2, equine RNA; lane 3, human RNA; these lanes were probed with full-length link protein cDNA. Lane 4, human RNA probed with a 400 bp fragment from the 3'-end of the 3' untranslated region of link protein. Size markers used were an ethidium bromide-stained RNA ladder (Gibco-BRL).

band failed to show the 528 bp fragment (results not shown). This larger mRNA species may reflect a rare transcript product synthesized by equine, porcine and rat chondrocytes, but not human chondrocytes. Interestingly, the G1 domain of aggrecan, which is similar in sequence and structure to link protein and is thought to have arisen by gene duplication, does not contain any comparable alternatively spliced variant in the rat [29] or human [7] aggrecan cDNA sequence.

Human chondrocyte RNA examined by Northern-blot analysis (Figure 3) showed three transcripts of 5.0, 2.9 and 1.55 kb in size when probed with full-length cDNA. Hybridization to the human RNA with a cDNA probe specific to the 3'-end of the 3' untranslated region detected only the 5.0 kb transcript, showing that splicing occurs in the 3' untranslated region which leaves the protein-coding region unchanged. Multiple transcripts of similar size are detected in porcine, equine (Figure 3) and bovine chondrocytes (results not shown), and in RNA derived from chicken chondrocytes [22] or rat chondrosarcoma cells [25].

Transcriptional initiation site mapping

Primer extension and S1 nuclease mapping were used to locate the transcription initiation start site. The relative positions of the primers that were used are shown in Figure 1. In primer extension experiments (Figure 4a) the primer 90/H resulted in a strong signal for a product of 130 bases. RNA prepared from all ages of cartilage specimens (10-53 years) and also from fetal cartilage RNA exhibited this band, which defined the transcription initiation at 315 bases upstream from the translation initiation codon. The S1 nuclease assays (Figure 4c) were carried out on the 20- and 26-year-old cartilage specimens which produced a band resolving at 130 bases. This length of protected probe, beginning at the 90/H primer position, confirmed initiation at 315 bases upstream from the translation initiation codon. Evi-

Figure 4 Primer extension and S1 nuclease analysis of link protein mRNA

(a) Primer extension with primer 90/H. (b) Primer extension with primer 91/B. (c) S1 nuclease analysis. M denotes sequencing ladders used as size markers and L929 a control experiment with mouse cell line RNA (no cross-hybridization to link protein). The arrows mark the extension or protected fragments described in the text in bp.

dence for the same initiation start site at 315 bases was also obtained when RNA prepared from the CaCo-2 cell line was investigated by primer extension and S1 nuclease mapping.

It has been reported [24] from primer extension analysis of human fetal cartilage RNA that the initiation start position was at a more 3' position than was shown by our results. This would predict an mRNA with a shorter, 122-base untranslated region, although results of nuclease protection were not reported in that study. We therefore carried out additional primer extension assays using a 35-base oligonucleotide primer (91/B) identical with that reported in the previous study in order to repeat the result. In Figure 4(b) all samples tested gave a single 240 base fragment (10-, 26-, 27-year-old, fetal and CaCo-2) which placed the transcription start site 315 bases upstream of the translation start. From these results the size of the first exon was confirmed to be 289 bp, which contrasts with the small 96 bp exon indicated by the earlier work. Our data is more analogous to the long 500 bp 5' untranslated region of the chicken link protein gene. Since we did not detect the short transcript in adult cartilage it is possibly a special form associated with early development. However, an investigation of a 26-week-old fetal cartilage chondrocyte RNA sample showed that it also lacked the short form of the mRNA. This makes it unlikely that there is significant usage of another start site in human chondrocytes during growth and development from the late fetus to the fully mature adult.

5'-flanking region

Analysis of the 5'-flanking region (Figure 1) revealed that the promoter does not have a consensus TATA box in close proximity (−35 to −25) to the transcription start. However, there is a TATA motif-like sequence, TCTAA, present at position −75. In the chicken link protein gene a consensus TATAA box was present at a similar distance [23] and it was suggested that this could function as a promoter in a transient expression assay. Interestingly, the chicken gene exhibits multiple start points within a 32 base region, a feature often found in genes lacking a classical promoter. Another feature of the human gene is a dinucleotide repeat (CA)₄CT(CA)₁₂ that lies within the transcription unit. Such repeats show polymorphisms in genes [30],

and they may have a role in transcriptional regulation of link protein expression.

This study was supported by the Arthritis and Rheumatism Council (U.K.) and the Medical Research Council (U.K.). The technical assistance of Alka Vaid is greatly appreciated.

REFERENCES

- Hardingham, T. E. and Muir, H. (1972) *Biochim. Biophys. Acta* **279**, 401–405
- Hardingham, T. E. and Fosang, A. J. (1992) *FASEB J.* **6**, 861–870
- Hardingham, T. E. (1979) *Biochem. J.* **177**, 237–247
- Heinegard, D. and Hascall, V. C. (1974) *J. Biol. Chem.* **249**, 4250–4256
- Périn, J. P., Bonnet, F., Thuriereau, C. and Jollès, P. (1987) *J. Biol. Chem.* **262**, 13269–13272
- Goetinck, P. F., Stirpe, N. S., Tsonis, P. A. and Carlone, D. (1987) *J. Cell Biol.* **105**, 2403–2408
- Doege, K. J., Sasaki, M., Kimura, T. and Yamada, Y. (1991) *J. Biol. Chem.* **266**, 894–902
- Perkins, S. J., Nealis, A. S., Dudhia, J. and Hardingham, T. E. (1989) *J. Mol. Biol.* **206**, 737–753
- Hardingham, T. E., Fosang, A. J. and Dudhia, J. (1994) *Eur. J. Clin. Chem. Clin. Biochem.* **32**, 249–257
- Mallein-Gerin, F., Koshier, R. A., Upholt, W. B. and Tanzer, M. L. (1988) *Dev. Biol.* **126**, 337–345
- Tsonis, P. A. and Goetinck, P. F. (1988) *Exp. Eye Res.* **46**, 753–764
- Fife, R. S., Caterson, B. and Myers, S. L. (1985) *J. Cell Biol.* **100**, 1050–1055
- Dudhia, J. and Hardingham, T. E. (1989) *J. Mol. Biol.* **206**, 749–753
- Dudhia, J. and Hardingham, T. E. (1990) *Nucleic Acids Res.* **18**, 1292
- Le Glédic, S., Périn, J. P., Bonnet, F. and Jollès, P. (1983) *J. Biol. Chem.* **258**, 14759–14761
- De Luca, S., Heinegard, D. and Hascall, V. C. (1977) *J. Biol. Chem.* **252**, 6600–6608
- Nguyen, Q., Murphy, G., Roughley, P. J. and Mort, J. S. (1989) *Biochem. J.* **259**, 61–67
- Nguyen, Q., Liu, J., Roughley, P. J. and Mort, J. S. (1991) *Biochem. J.* **278**, 143–147
- Martin, H. and Dean, M. F. (1993) *Eur. J. Biochem.* **212**, 87–94
- Liu, J., Cassidy, J. D., Allan, A., Neame, P. J., Mort, J. S. and Roughley, P. J. (1992) *J. Orthop. Res.* **10**, 621–630
- Mort, J. S., Poole, A. R. and Roughley, P. J. (1983) *Biochem. J.* **214**, 269–272
- Deák, F., Kiss, I., Sparks, K. J., Argraves, W. S., Hampikian, G. and Goetinck, P. F. (1986) *Proc. Natl. Acad. Sci. U.S.A.* **83**, 3766–3770
- Deák, F., Barta, E., Mestric, S., Biesold, M. and Kiss, I. (1991) *Nucleic Acids Res.* **19**, 4983–4990
- Rhodes, C., Savagner, P., Line, S., Sasaki, M., Chirigos, M., Doege, K. and Yamada, Y. (1991) *Nucleic Acids Res.* **19**, 1933–1939

-
- 25 Rhodes, C., Doege, K., Sasaki, M. and Yamada, Y. (1988) *J. Biol. Chem.* **263**, 6063–6067
- 26 Watt, F. M. and Dudhia, J. (1988) *Differentiation* **38**, 140–147
- 27 Chomczynski, P. and Sacchi, N. (1987) *Anal. Biochem.* **162**, 156–159
- 28 Fogh, J., Fogh, J. M. and Orfeo, T. (1977) *J. Natl. Cancer Inst.* **59**, 221–226
- 29 Doege, K., Sasaki, M., Horigan, E., Hassell, J. R. and Yamada, Y. (1987) *J. Biol. Chem.* **262**, 17757–17767
- 30 Oudet, C., Heilig, R. and Mandel, J. L. (1990) *Human Genet.* **84**, 283–285
- 31 Sharp, P. A. (1981) *Cell* **23**, 643–646
- 32 Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472

Received 7 March 1994/18 April 1994; accepted 28 April 1994