Volume 311, number 3, 305–310 FEBS 11685 October 1992

Genomic organization and full-length cDNA sequence of human collagen X

Ernst Reichenberger[a], Frank Beier[a], Phyllis LuValle[b], Bjorn R. Olsen[b], Klaus von der Mark[a] and Wolf M. Bertling[a]

[a]*Max-Planck Society, Clinical Research Unit for Rheumatology at the Medical Clinic III, University of Erlangen-Nürnberg, Schwabachanlage 10, 8520 Erlangen, Germany* and [b]*Department of Anatomy and Cellular Biology, Harvard Medical School, Boston, MA 02115, USA*

Received 7 September 1992

We have determined the full-length cDNA sequence of the human α1(X) collagen gene by sequence analysis of a genomic clone ERG [(1991) Dev. Biol. 148, 562–572], and of cDNA fragments generated from a reverse transcribed as α1(X) mRNA by PCR. We defined the promoter region, the transcription initiation site and the full-length 5′-untranslated region. We also report the exon/intron boundaries of the transcript and the complete 3′-untranslated region as well as a 3′-flanking sequence containing two additional polyadenylation signals. The promoter region is homologous to chicken and mouse type X promoters within several highly conserved regions. The genomic organization shows high homologies to chicken and mouse.

Human cartilage; Collagen; Human type X; Nucleotide sequence; Regulatory sequence; Evolution

1. INTRODUCTION

Collagen type X is a developmentally expressed cartilage specific component of the extracellular matrix (ECM) which in healthy cartilage is restricted to zones of endochondral ossification [2–5]. It is expressed by hypertrophic chondrocytes in the calcifying zone of the growth plate [1,6], in zones of secondary ossification [1], in newly forming osteophytic or reparative cartilage [7,8], or in bone callus during fracture healing [9]. This restricted expression to zones of degrading cartilage implies that collagen type X may play an important role in the regulation of endochondral ossification and cartilage degradation.

Collagen type VIII and type X have similar structural characteristics, such as a triple-helical domain, half as long as in fibrillar collagens, in between a short N-terminal and a larger C-terminal globule [10–16] and are therefore classified as short collagens [4]. There is a remarkable sequence homology between collagen type VIII and X among species, and both collagens are able to assemble in a hexagonal lattice structure [17,21].

The collagen type X gene, as well as the type VIII genes are distinct from all other collagens in their condensed gene structure [14–16,19,20]. Collagen type VIII is a heterotrimeric protein consisting of α1 and α2 chains, whereas collagen type X is a homotrimer. The triple-helical domains are interrupted by several imperfections in the G-X-Y triplet structure. In the case of type X collagen these interruptions are well conserved between chicken, bovine, mouse, and human.

Although much of our current understanding about the structure and function of collagen type X is derived from the avian system [21], detailed analysis of its regulation is needed also in mammalian systems, due to the differences in the cartilage differentiation and bone formation patterns in birds and mammals. In the chicken long bone, endochondral ossification occurs in an irregular pattern along the entire diaphysis [22,23]. In mammals, however, it occurs in a narrow zone perpendicular to the long bone axis from the mid-diaphysis to the epiphysis. Therefore the zone of hypertrophic cartilage in chicken long bones is considerably larger than in mammals.

We have recently published studies on the developmental expression in fetal cartilage, in adult healthy, as well as in osteoarthritic and rheumatoid adult cartilage [1,7,8] indicating precisely regulated gene activation. Changes of in vitro expression patterns of type X collagen due to the influences of retinoic acid [24], of vitamin D [25] and of calcium β-glycerophosphate [26] also suggest a precise and highly complex regulation of the expression of its gene.

To enable further studies of the regulation of the collagen type X gene in a mammalian system we defined the complete structure of human type X collagen and began to study the enhancer and promoter regions. This

Correspondence address: W.M. Bertling, Max-Planck Society, Clinical Research Unit for Rheumatology at the Medical Clinic III, University of Erlangen-Nürnberg, Schwabachanlage 10, 8520 Erlangen, Germany. Fax: (49) (9131) 20 6951.

system allows investigation of pathologic deficiencies, such as in rachitic growth cartilage or chondrodysplasias.

2. EXPERIMENTAL

A 329 bp fragment (pERX) coding for parts of the C-terminal globular domain was generated by PCR from human genomic DNA [1]. This fragment was initially used to screen a human genomic library (HL 1067J; Clontech, CA, USA). Hybridization of the library was performed as described [27] in 50% formamide at 42°C, washing was done twice with 2 × SSC for 5 min, twice with 2 × SSC, 0.1 × SDS at 65°C for 15 min and with 0.1 × SSC at room temperature for 10 min. Of 300,000 screened clones, four hybridized with our probe. A λ clone (ERG) with a 13 kb *Xho*I insert (*Xho*I cuts the insert into two fragments of about 6700 bp) (Fig. 1) was identified and used for further analysis after recloning both *Xho*I-fragments into pBluescript SK$^+$ (Stratagene). This λ clone covers the locus of collagen type X from about 2800 bp upstream of the transcription initiation site to about 3800 bp downstream of the polyadenylation site.

Fragments covering the coding sequence were identified by Southern analysis. Sequencing of both strands [28] using Sequenase 2.0 (USB, CA, USA) and ^{35}S-dATP. Double-strand and single-strand sequencing was carried out with either vector specific or sequence specific oligonucleotides generated on a oligonucleotide synthesizer (Gene Assembler Plus, Pharmacia, Sweden). Exon sequences were analyzed by sequence comparison with published chicken [19] and bovine [13] sequences using the computer program Genpro version 5 (Riverside Scientific Enterprises, WA, USA).

The transcription initiation site was identified by S1-mapping and primer extension assays [30] (Fig. 3). For all the following experiments total RNA prepared from the same batch of freshly isolated fetal growth-plate-chondrocytes (24 weeks of gestation) was used. Hybridization of the primers to 10 μg of the total RNA was performed in 50% formamide and 0.4 M NaCl overnight in a volume of 30 μl at temperatures calculated by the primer analysis program Oligo (MedProbe A.S., Norway). S1-nuclease reactions were performed in a final volume of 330 μl with 200 units S1 nuclease (Boehringer-Mannheim, Germany) with a single-stranded 569 bp antisense probe created by primer extension of a clone comprising the ligation product of exon 2, exon 1 and 460 bp of the 5'-flanking region with the downstream primer Ext5 (5'-GCACGCAGAATCCATCTGAGAA-TATGCTGCCACAAATACC-3'). For primer extension assays the reverse transcription was performed in a final volume of 20 μl with the same primer as used for S1 analysis (Ext5). Exon/intron boundaries were identified by consensus sequences and verified by polymerase chain reaction (PCR) of cDNA with primers comprising presumed exon sequences. PCR was done according to [31] in a thermocycler (PREM III, LEP Scientific, GB).

3. RESULTS AND DISCUSSION

The third exon which comprises the C-terminal globular domain and the triple-helical region was identified by Southern hybridization. We located exon 3 with our cDNA probe (pERX) which we had obtained by PCR [1] on a 1.4 kb *Xho*I/*Hin*dIII fragment (Fig. 1). Sequencing of this and neighboring fragments revealed an open reading frame of 2195 nucleotides. 1036 nucleotides further downstream the first polyadenylation site was found. Two more polyadenylation sites were detected on the next 100 nucleotides (Fig. 2).

To define the 3'-end of the transcript and the active polyadenylation site we performed reverse transcription

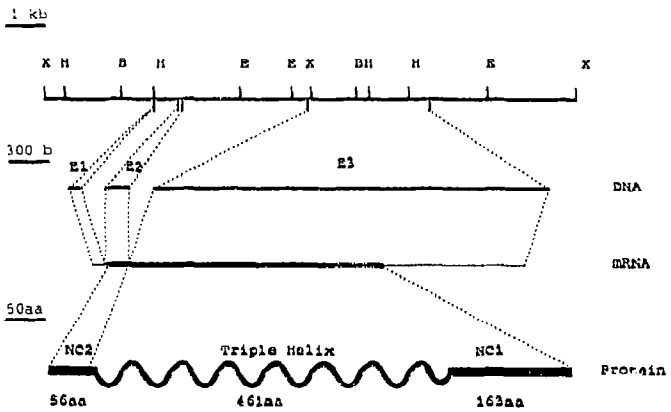

Fig. 1. Diagram showing a restriction map of the human genomic clone ERG (upper part) and the gene structure of human collagen type X. The genomic clone was mapped with the restriction enzymes *Xho*I (X), *Hin*dIII (H), *Bam*HI (B), and *Eco*RI (E). The intron–exon splice junctions of the exons 1–3 within the mRNA are shown in relation to the genomic clone. The protein structure is indicated as deduced from the nucleotide sequence (NC2, NC1: non-collagenous globular domains).

with an oligo(dT) primer and total RNA as a template followed by PCR with the oligo(dT) primer and an upstream primer ERO61 (5'-TGAAATTTGATTTGA-GAAACTCGGC-3'). The cloning and sequencing of this amplification product identified the first polyadenylation site (Fig. 2) as the preferred site. A second polyadenylation site has been reported by [14] 94 nucleotides downstream as part of a 34 bp direct repeat (corresponding to nucleotides 3259–3293 in Fig. 2). In our sequence we could not confirm this finding, but instead detected a third polyadenylation signal (Fig. 2: nucleotide 3277–3283). We verified our sequence derived from the human genomic clone ERG by PCR of total human genomic DNA as template with ERO61 as upstream primer and a primer specific for our divergent sequence, ERO65 (5'-TTTATTGTCCTACTTTTTTATTAAC-3'). Using an *Eco*RI-linked downstream oligonucleotide ERO63 (5'-CCGAATTC TTGAGAACAGCAAATT-GCTG-3') priming in the 3'-flanking region between the first and second polyadenylation site and ERO61 as upstream primer we could also show that neither of the aditional polyadenylation signals are utilized in this mRNA from freshly isolated chondrocytes.

A dA-rich region 324 nucleotides downstream of the translation stop in the 3'-untranslated region (UTR) might explain why we initially had difficulties in obtaining correctly sized PCR amplification products of cDNA from RNA of hypertrophic chondrocytes, since our oligo(dT) primers would anneal to this region as well. Although this dA-rich region (47 of 50 bases are A) is not as well conserved in other species it might be a remnant of a former poly(A) tail. If the collagen type X gene arose by reintegration of a processed collagen transcript, as had been suspected before [19], it would

Fig. 2. Nucleotide sequence of the full-length cDNA, 5′-flanking as well as 3′-flanking regions, and of exon–intron splice junctions. Numbering of the sequence refers to the cDNA sequence and does not count intron sequences. Vertical arrows mark the description start and termination sites. The transcription start site has been designated +1. Consensus sequences for exon–intron splice junctions and pyrimidine-rich sequences at splice acceptor sites are overlined. Intron sequences are printed in lower case letters. The presumed TATA-box, and polyadenylation sites are underlined. Dots indicate deviations from the previously published sequences [14,15] and nucleotides number 3259–3293 deviate from the sequence of Thomas et al. [14]. Wavy lines indicate additional but unused splice acceptor sites and polyadenylation sites. Horizontal arrowheads indicate the start of the triple-helix and the NH_2- and COOH-terminal globular domains.

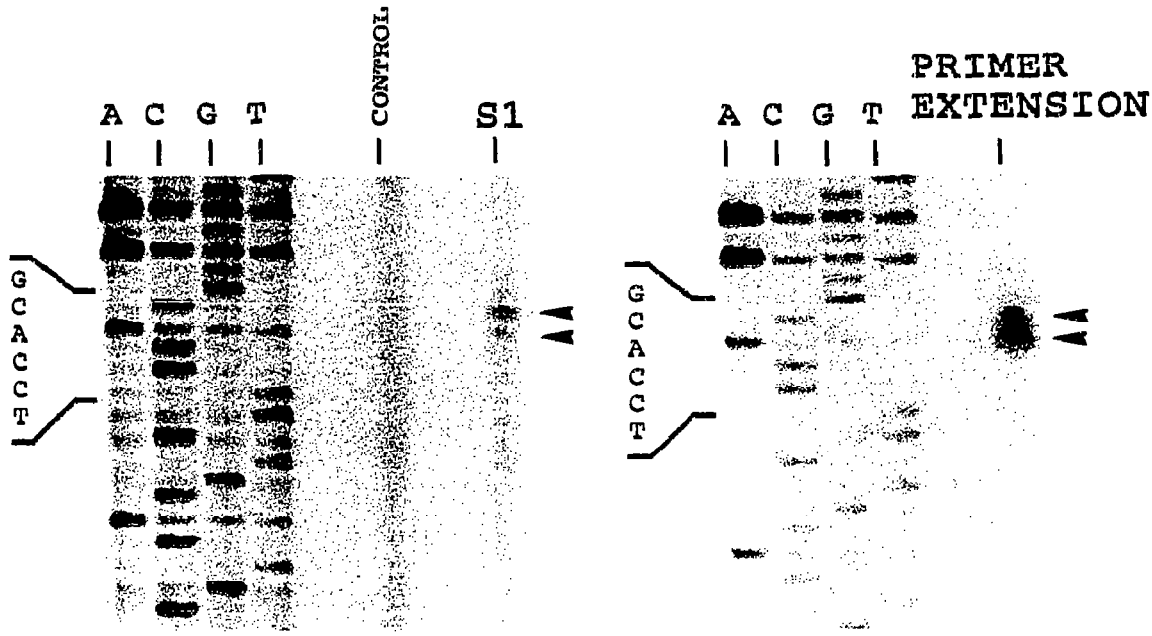

Fig. 3. The transcription start site was determined by S1-analysis (left panel) and by primer extension analysis (right panel). The same primer (Ext5) was used for the synthesis of a single-strand antisense fragment for S1-analysis, for primer extension and for sequencing reactions (A C G T lanes). Left of the panels, a part of the sequence is shown as read from the gel. As a control for S1-analysis the same amount of the 569 bp antisense fragment was applied after incubation without RNA and after digestion with S1-nuclease under identical conditions. Arrowheads indicate the transcription start sites.

```
hum AACCATAATTAGTGCTTACCAACTTTGAGGGGAGGAGCTTAAGCTCAG-GGTAACCTCATGTAGTAAGGG--AAAA -36
chk TCAGATAATTACTGCTTACTCTCT-TG--CCGAGGAGCTTATGTTCAGCAGTCACC----AAACT-AGGGTGAAAA

hum ACAGTATAAATA-CCACTGAACAGCTCTTGAGA--GGCACCTTCTGCACTGCTCATCTG-GGCAGAGGAA-GCTTC  37
chk TC-GTATAAATAGTCA-AGGGTGACGCTTGAAATCATCAGCTTCTGCTC-ACTCACCAGTGGCAGAAGAACTCCTC
```

Fig. 4. Comparison between the promoter region of human (upper sequence) and chicken (lower sequence). Transcription start sites are marked by arrows. Asterisks (*) mark identical nucleotides, colons (:) stand for conserved purines and pyrimidines. A conserved TATA-box is marked by lines.

explain why 92% of the 3189 bp long mRNA are encoded by only one exon (exon 3) although other collagen genes have abundant introns [32–34]. One would have to require that the integration happened into the gene locus of a different gene before type VIII and type X genes developed separately. The integration of a reverse transcript into another gene, yielding a new functional genetic entity, is certainly a rare event. A recombinational event later in evolution might, however, have contributed to conserve the promoter part of this gene between such distantly related species as chicken and man (see below and Fig. 4). Browsing through the sequence of intron 2 we found indications that the target of this reintegration process might have been another collagen gene. 20 nucleotides upstream of the actual splice acceptor site of exon 3 there is a second non-utilized splicing signal. Both unused sites, the splice acceptor site and the lariat forming consensus site in intron 2, are fairly well conserved [35] (Fig. 2) and upstream of this sequence in intron 2 we find remnants of a reading frame coding for several G-X-Y entities.

Sequences upstream of the internal *Xho*I site (Fig. 1) were searched for open reading frames (ORFs) which showed homology to coding regions of known collagen type X sequences. About 3 kb upstream of exon 3 we detected a sequence with high homology to exon 2 of chicken and, as we learned during the course of our studies, to bovine [13] and mouse [16] exon 2 (Figs. 1 and 2). This designated exon 2 in human comprises 169 nucleotides of which 154 bases are translated and 15 belong to the 5′-UTR. We verified the use of this exon sequence on the RNA level by PCR. The published sequences for exon 2 and 3 [14,15] were in agreement with our sequence except exchanges of single nucleotides as indicated in Fig. 2 and the described deviation in the 3′-flanking region.

Sequencing further upstream of exon 2 we discovered several fairly well conserved splice donor sites [36] in a

distance of 0.6–1.2 kb. We then decided to test for the usage of these sites by PCR. We used RNA of freshly isolated hypertrophic chondrocytes to generate a cDNA with a primer from within exon 3, ERO50 (5'-CTGGTTTCCCTACAGCTGA-3'), and used this cDNA for PCR with primers situated about 10 bp 5' of splice donor consensus sequences. With one of these upstream primers (5'GAGGAAGCTTCAGAAAGC-TG) and ERO50 as downstream primer we amplified a fragment which we then cloned and sequenced. This sequence defined the exon boundaries of exon 3, 2, and 1.

The transcription start site was defined by primer extension analysis as 81 bp upstream of the determined splice donor site for exon 1 (Fig. 3, panel B). This result was verified by S1-analysis (Fig. 3, panel A). The preferred transcription initiation was defined as the nucleotide C followed by the nucleotid A in the sequence 5'GAGGCACC (Figs. 2 and 3). This transcription site shows a similarity to the inverted consensus sequence GYCTC [37]. Human promoter sequences and portions of the 5'-UTR are highly homologous to corresponding chicken sequences. Other processing signals, however, such as splice donor and acceptor sites do not show such a high degree of homology. The sequence homology among human and chicken between the transcription initiation and the TATA-box is lower than in surrounding stretches. This might be due to differently sized polymerase II complexes in chicken and man, which prompted a selection for different distances between the pre- and post-transcription initiation boxes. Another remarkable feature is a homologous region upstream of the TATA-box, a direct repeat (5'-GWGCTTA) (−96 to −92 and −74 to −68). A part of this sequence (nucleotide number −83 to −72 (Fig. 2) resembles the binding motif of transcription factors of the Ets family of protooncogenes which bind to a purine-rich motif around a conserved GAG trinucleotide [38]. AT-rich motifs at a distance of about 100 nucleotides to the transcription initiation site have also been found in human [39] and rat [40] and identified as binding sites for HNF-1 respectively Pit-1. The sequence of the mouse promoter region [16] is also strikingly similar to the human sequence. One of the few stretches of reduced homology is in fact the actual transcription initiation site, surrounded upstream and downstream by regions of high conservation. While the homology to the promoter region of chicken ends rather abruptly around position −103, it seems not to be accidental, that upstream of this position the A/T-content rises. Such an increase has also been observed for a second mammalian collagen type X gene [16].

This complexly regulated gene with its highly conserved promoter sequences and exon/intron organization will be further analyzed in order to contribute to our understanding of developmental and tissue specific regulation.

Acknowledgements: We thank C. Matzner for excellent technical assistence. This work was supported by the BMFT (Grant 01VM87020 to K.v.d.M.; Grant 01VM8819/9 to W.M.B.) and the NIH (Grant AR36819 to B.R.O.).

REFERENCES

[1] Reichenberger, E., Aigner, T., von der Mark, K., Stöss, H. and Berling, W. (1991) Dev. Biol. 148, 562–572.
[2] Schmid, T.M. and Conrad, H.E. (1982) J. Biol. Chem. 257, 12451–12457.
[3] Gibson, G.J. and Flint, M.H. (1985) J. Cell Biol. 101, 277–284.
[4] Schmid, T.M. and Linsenmayer, T.F. (1987) in: Biology of the Extracellular Matrix: a series; Structure and Function of Collagen Types (Mayne, R. and Burgeson, R.E. eds.) pp. 223–258, Academic Press, New York.
[5] Mayne, R. (1989) Arthritis Rheum. 32, 241–246.
[6] Kirsch, T. and von der Mark, K. (1991) Eur. J. Biochem. 196, 575–580.
[7] Aigner, T., Reichenberger, E., Bertling, W.M., Stoess, H. and Von der Mark, K. (1992) submitted.
[8] Von der Mark, K., Kirsch, T., Nerlich, A., Kuß, A., Weseloh, G., Glückert, K. and Stöß, H. (1992) Arthritis Rheum. 35, 806–811
[9] Grant, W.T., Wang, G.J. and Balian, G. (1987) J. Biol. Chem. 262, 9844–9849.
[10] Yamaguchi, N., Benya, P.D., van der Rest, M. and Ninomiya, Y. (1989) J. Biol. Chem. 264, 16022–16029.
[11] Muragaki, Y., Mattei, M.G., Yamaguchi, N., Olsen, B.R. and Ninomiya, Y. (1991) Eur. J. Biochem. 197, 615–622.
[12] Ninomiya, Y., Gordon, M., van der Rest, M., Schmid, T., Linsenmayer, T. and Olsen, B.R. (1986) J. Biol. Chem. 261, 5041–5050.
[13] Thomas, J.T., Kwan, A.P., Grant, M.E. and Boot-Handford, R.P. (1991) Biochem. J. 273, 141–148.
[14] Thomas, J.T., Cresswell, C.J., Rash, B., Nicolai, H., Jones, T., Solomon, E., Grant, M.E. and Boot-Handford, R.P. (1991) Biochem. J. 280, 617–623.
[15] Apte, S.S., Seldin, M.F., Hayashi, M. and Olsen, B.R. (1992) Eur. J. Biochem. 206, 217–224.
[16] Elima, K., Eerola, I., Rosati, R., Metseranta, M., Garofalo, S., Perälä, M., de Crombrugghe, B. and Vuorio, E. (1992) Abstract: 13. FECTS Meeting, Davos
[17] Sawada, H., Konomi, H. and Hirosawa, K. (1990) J. Cell Biol. 110, 219–227.
[18] Chen, Q., Linsenmayer, C., Gu, H., Schmid, T.M. and Linsenmayer, T.F. (1992) J. Cell Biol. 117, 687–694.
[19] LuValle, P., Ninomiya, Y., Rosenblum, N.D. and Olsen, B.R. (1988) J. Biol. Chem. 263, 18378–18385.
[20] Yamaguchi, N., Mayne, R. and Ninomiya, Y. (1991) J. Biol. Chem. 266, 4508–4513.
[21] Lutfi, A.M. (1971) Acta Anat. Basel 79, 27–35.
[22] Von der Mark, K., von der Mark, H. and Gay, S. (1976) Dev. Biol. 53, 153–170.
[23] Descalzi-Cancedda, F., Gentili, C., Manduca, P. and Cancedda, R. (1992) J. Cell Biol. 117, 427–435.
[24] Kwan, A.P., Dickson, I.R., Freemont, A.J. and Grant, M.E. (1989) J. Cell Biol. 109, 1849–1856.
[25] Thomas, J.T., Boot-Handford, R.P. and Grant, M.E. (1990) J. Cell Sci. 95, 639–648.
[26] Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1988) Current Protocols in Molecular Biology, Greene Publishing Ass./Wiley-Interscience, New York.
[27] Sanger, F. and Coulson, A.R. (1978) FEBS Lett. 87, 107–110.
[28] Maniatis, T., Fritsch, E.F. and Sambrook, J., (1988) Molecular cloning: a laboratory manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

[29] Krawetz, S.A., Pon, R.T. and Dixon, G.H. (1989) Nucleic Acids Res. 17, 819.
[30] Vuorio, E. and de Crombrugghe, B. (1990) Annu. Rev. Biochem. 59, 837–872.
[31] van der Rest, M. and Garrone, R. (1991) FASEB J. 5, 2814–2823.
[32] Sandell, L.J. and Boyd, C.D. (1990) in: Extracellular Matrix Genes (Sandell, L.J. and Boyd, C.D. Eds.) Academic Press, New York.
[33] Sharp, P.A. (1987) Science 235, 766–771.
[34] Cech, T.R. (1986) Cell 44, 207–210.
[35] Seto, E., Shi, Y. and Shenk, T. (1991) Nature 354, 241–245.
[36] Wang, C.Y., Petryniak, B., Ho, I.C., Thompson, C.B. and Leiden, J.M. (1992) J. Exp. Med. 175, 1391–1399.
[37] Li, Y., Shen, R.F., Tsai, S.Y. and Woo, S.L. (1988) Mol. Cell Biol. 8, 4362–4369.
[38] Resendez Jr., E., Wooden, S.K. and Lee, A.S. (1988) Mol. Cell. Biol. 8, 4579–4584.