The mouse collagen X gene: complete nucleotide sequence, exon structure and expression pattern

Kati ELIMA,* Iiro EEROLA,* Rita ROSATI,† Marjo METSÄRANTA,* Silvio GAROFALO,† Merja PERÄLÄ,* Benoit DE CROMBRUGGHE† and Eero VUORIO*‡

*Departments of Medical Biochemistry and Molecular Biology, University of Turku, Kiinamyllynkatu 10, 20520 Turku, Finland, and †Department of Molecular Genetics, The University of Texas M. D. Anderson Cancer Center, Houston, TX, U.S.A.

Overlapping genomic clones covering the 7.2 kb mouse $\alpha 1(X)$ collagen gene, 0.86 kb of promoter and 1.25 kb of 3'-flanking sequences were isolated from two genomic libraries and characterized by nucleotide sequencing. Typical features of the gene include a unique three-exon structure, similar to that in the chick gene, with the entire triple-helical domain of 463 amino acids coded by a single large exon. The highest degree of amino acid and nucleotide sequence conservation was seen in the coding region for the collagenous and C-terminal non-collagenous domains between the mouse and known chick, bovine and human collagen type X sequences. More divergence between the sequences occurred in the N-terminal non-collagenous domain. Similarity between the mammalian collagen

X sequences extended into the 3'-untranslated sequence, particularly near the polyadenylation site. The promoter of the mouse collagen X gene was found to contain two TATAA boxes 159 bp apart; primer extension analyses of the transcription start site revealed that both were functional. The promoter has an unusual structure with a very low G + C content of 28 % between positions -220 and -1 of the upstream transcription start site. Northern and *in situ* hybridization analyses confirmed that the expression of the $\alpha 1(X)$ collagen gene is restricted to hypertrophic chondrocytes in tissues undergoing endochondral calcification. The detailed sequence information of the gene is useful for studies on the promoter activity of the gene and for generation of transgenic mice.

INTRODUCTION

Collagen X has several unique features among the collagen gene and protein family. It exhibits a very restricted expression pattern confined to one developmental stage of endochondral ossification, i.e. the zone of hypertrophic chondrocytes [1,2]. These cells represent the terminally differentiated form of chondrocytes. Before hypertrophy these cells divide and produce a cartilaginous matrix, and are thus responsible for the growth of long bones at epiphyseal growth plates [3,4]. Collagen X appears in the cartilage matrix shortly before its mineralization and replacement by bone [1,2]. Experiments *in vivo* and in cell culture have shown that upon hypertrophy the chondrocytes undergo dramatic changes in their gene expression, including a switch from type II to type X as the major collagen type produced [2,5,6].

The structure of the collagen X gene is also unique in the collagen family. Unlike the other collagen genes, which all have a characteristic multiexon structure coding for the collagenous domains [7], the triple-helical domain of collagen X in the chick is coded for by a single exon of over 2000 bp. Although the complete gene structure is currently known only for the chick [8,9], all available evidence suggests that the bovine and human collagen X genes also share the same condensed structure [10,11]. Recently the genes for rabbit, mouse and human $\alpha 1$ and $\alpha 2$ chains of collagen VIII were placed in the same subfamily of short chain collagens, as they were shown to share a genomic structure very similar to that of collagen X [12,13].

Another unique feature of collagen X is its unusually high melting temperature of 47 °C [2,14]. In spite of the short triple-helical domain with eight interruptions, the thermal stability of

collagen X is greater than that of fibrillar collagens. The molecule has two cleavage sites for vertebrate collagenase [15], but the major cleavage product of 32 kDa still has a melting temperature of 43 °C and thus remains triple-helical at physiological temperatures. The collagen X molecule is a homotrimer of three 59 kDa $\alpha 1(X)$ chains. Except for the removal of the signal peptide there appears to be no post-translational proteolytic processing of this collagen type [2]. Two supramolecular forms have been suggested for collagen X in embryonic chick cartilage [16]. By immunoelectron microscopy the molecule has been localized both in fine filamentous mats around hypertrophic chondrocytes and in association with collagen II fibrils. *In vitro*, collagen X has been shown to associate into multimeric clusters forming a hexagonal lattice analogous to that seen in Descemet's membrane [17]. Other experiments have shown that collagen X can migrate through a cartilaginous matrix and bind to collagen II fibrils [18].

Despite all the recent information on the primary and supra-molecular structure of collagen X, its genomic organization and its expression by hypertrophic chondrocytes, very little is known about its function(s) in the hypertrophic cartilage matrix. The very restricted tissue distribution of collagen X in cartilages committed to calcification and replacement by bone suggests a function in the process of calcification and/or matrix removal [2]. No mutations of the collagen X gene which disrupt the normal function of the protein are known. Introduction of specific mutations into the gene and expressing the mutant protein in transgenic mice should therefore be a useful approach to learn more about the function of collagen X. In preparation for such experiments, we report here the isolation and detailed characterization of the mouse $\alpha 1(X)$ collagen gene.

Abbreviation used: 1 × SSC, 0.15 M NaCl/0.015 M trisodium citrate, pH 7.0. In this paper, the collagen types are written as, e.g., 'collagen X', instead of the style 'type X collagen' which has been used in other publications by these authors.

‡ To whom correspondence should be addressed.

The nucleotide sequence data reported in this paper have been submitted to the GenBank/EMBL DataBank with accession number X67348.

EXPERIMENTAL

Genomic libraries

Two commercial mouse genomic libraries were screened before the full gene sequence of the murine collagen X was obtained; one was a cosmid library prepared with liver DNA from adult Balb/c mice (Clontech), and the other was a phage library prepared with DNA from NIH-3T3 cells (Stratagene). The genomic library in cosmid pWE15 vector was screened first using a [³²P]dCTP-labelled 721 bp cDNA fragment of mouse $\alpha 1(X)$ collagen clone pMCol10a1-1 [19] as probe. The hybridization was performed in 5 × SSC, 10 × Denhardt's solution, 0.5 % SDS and 200 µg/ml denatured calf thymus DNA at 65 °C overnight. After the high-stringency washes (2 × SSC/0.5 % SDS at room temperature, and 0.2 × SSC/0.5 % SDS at 65 °C for 1 h with three changes) and autoradiography, two positive clones were identified among the 5 × 10⁵ clones screened.

The genomic library in λ FIX II vector was screened with a 1.2 kb *Bam*HI fragment corresponding to the 5'-end of cosmid clone cKEX-8. Hybridization as above yielded five phage clones covering the 5'-end of the gene (λ RR-1 to -5). Purified genomic clones were characterized by restriction mapping, Southern hybridization and nucleotide sequencing.

Subcloning and sequencing

For sequencing, approx. 9.3 kb of the gene was first subcloned as 34 subclones in Bluescript SK⁻ vector (Stratagene). In addition to oligonucleotide primers corresponding to the T3 and T7 recognition sites of the vector, 20 synthetic oligonucleotides were used as sequencing primers. Sequencing was performed on the double-stranded DNA using the Sequenase reagent kit and [³⁵S]dATP. The sequences were stored and analysed using the GCG software [20].

RNA extraction, cDNA synthesis and amplification by PCR

Total RNAs were isolated from several newborn mouse tissues including rib and epiphyseal cartilages using the guanidinium isothiocyanate method [21]. Maloney Murine Leukemia Virus (MMLV) reverse transcriptase was used for cDNA synthesis using total RNA (10 µg) as the template under conditions suggested by the supplier (Gene Amp RNA PCR kit, Perkin-Elmer Cetus) with random hexamers as primers. Aliquots of cDNA were used for amplification by the PCR using specific oligonucleotide primers IE-1 and IE-2, corresponding to sequences in exon 2 and in the putative exon 1 respectively (approximate locations shown in Figure 1). The reactions were cycled by denaturing at 94 °C for 1 min, annealing at 57 °C for 2 min, and extension at 72 °C for 3 min. After 30 amplification cycles, aliquots of the reaction products were fractionated by electrophoresis on 1.5 % agarose gels. The 167 bp fragment obtained was purified and cloned by blunt-end ligation into an *Eco*RV site of the Bluescript vector.

Primer extension

The start site of transcription of the mouse collagen X gene was analysed by primer extension. Total RNA (15 µg) from epiphyseal cartilages of newborn mice served as a template for reverse transcriptase, and two specific oligonucleotides B (5'-CCCTGGAGTATTACTGTTTCTCCCT-3') and C (5'-CTGGTGCTGGGAGATGCGTGGGGCAGCTTTC-3') labelled at the 5'-end as the primer for cDNA synthesis as

described earlier [22]. The products were fractionated on a sequencing gel in parallel with dideoxy sequencing reactions primed with the same oligonucleotides.

Northern and *in situ* hybridizations

Total RNAs isolated from several mouse tissues were fractionated on 0.7 % agarose gels under denaturing conditions and transferred on to Pall Biodyne membranes as suggested by the supplier (Pall Ultrafine Filtration Corp., Glen Cove, NY, U.S.A.). The hybridizations were performed at 42 °C in 50 % formamide, 5 × SSC, 2 × Denhardt's solution, 0.1 % SDS and 500 µg/ml sonicated calf thymus DNA, and the high-stringency washes were at 56 °C in 0.2 × SSC/0.1 % SDS.

In situ hybridizations were carried out as described previously [23]. In brief, entire 18-day mouse embryos or selected tissue samples were fixed with formalin and embedded in paraffin for sectioning. The sections for *in situ* hybridization were pretreated with proteinase K and HCl and acetylated. The hybridizations were carried out at 42 °C for 24 h using a [³⁵S]deoxy(thio)ATP-labelled 1.2 kb *Bam*HI fragment from exon 3 sequences as probe, followed by washing, autoradiography at 4 °C for 5–20 days, and staining of the sections with haematoxylin [23]. The mRNA for collagen II was detected similarly using the labelled insert from the corresponding cDNA clone pMCol2a1-1 [24] as the probe.

RESULTS AND DISCUSSION

Screening of the mouse genomic cosmid library with the cDNA probe pMCol10a1-1 for mouse $\alpha 1(X)$ collagen [19] yielded two positive clones. Restriction mapping and sequencing of these overlapping clones confirmed that they coded for the 3'-end of the mouse $\alpha 1(X)$ collagen gene. The same cosmid library was rescreened with the 1.2 kb *Bam*HI fragment from the 5'-end of clone cKEX-8 exhibiting the longest 5'-extension (Figure 1). Since no additional positive clones were obtained, a mouse genomic phage library was screened with the same probe. Five additional clones were isolated and identified as genomic clones for the mouse $\alpha 1(X)$ collagen. Clone λ RR-5 containing the longest 5'-extension was selected for restriction mapping and sequencing (Figure 1).

The complete nucleotide sequence of the mouse collagen X gene was determined from genomic (NIH-3T3) DNA in the phage clone λ RR-5, except where DNA from the Balb/c cosmid library was available from clone cKEX-8 (Figures 1 and 2). The gene spans 7.2 kb from the transcription start site to the polyadenylation signal and contains three exons. Exons were identified by flanking consensus splice signals and by comparison with the corresponding chick gene sequences [9,25]. The intron-exon boundaries between exons 1 and 2 were confirmed by cloning and sequencing of a cDNA clone covering the corresponding mRNA sequence using fragments amplified by PCR (see Figure 1).

Exon structure

The overall organization of the gene (Figures 1 and 2; Table 1) shows an unusual three-exon structure, similar to that described for chick collagen X [9]. Shown in Figure 2 is the coding sequence of the mouse $\alpha 1(X)$ collagen gene, including exon-intron boundaries, the promoter and 1.1 kb 3'-untranslated sequence of the transcript, compared with the corresponding bovine cDNA sequence [10]. Table 1 summarizes the current information of the exon-intron structure of the chick, bovine, human and mouse collagen X genes. A complete nucleotide sequence is known only

Figure 1 Restriction map of the mouse collagen X gene

Top, schematic presentation of the genomic clones λ RR-5 and cKEX-8 and the exon-intron structure of the gene, with the location of the coding region shown by the black boxed areas. Middle, restriction sites for nine enzymes with less than 10 recognition sites within the sequence. Bottom, the scale in kilobases.

for the mouse gene, but the information available suggests that all collagen X genes share a similar three-exon structure consisting of a short non-coding exon 1, intron 1 of approx. 600 bp, exon 2 of 159–169 bp, intron 2 of 2000–3400 bp, and a large exon 3 coding for the entire triple-helical domain, the C-terminal globular domain and a 3'-untranslated domain of approx. 1000 bp in the mammalian species [10,11]. In chick the 3'-untranslated domain is considerably shorter [9,25].

Promoter and transcription start site

Primer extension analysis was performed to identify the transcription start site of the gene (Figure 3). Since two TATAA sequences were seen in the putative promoter region which were 158 bp apart, two different oligonucleotide primers were used for primer extension. The analyses revealed that both TATAA boxes are functional, since two different transcription start sites were observed 44 bp and 31 bp from the upstream and downstream TATAA boxes respectively (Figures 2 and 3). The promoter contains one CCAAT element on the template strand at position -65 (-228 from the downstream transcription start site). Considerable divergence of the 5'-untranslated sequences is observed between mouse, bovine and chick genes. Interestingly, in chick the 5'-end of exon 1 and the TATAA box [9] show some sequence similarity with the downstream promoter sequence of the mouse gene, while the 5'-end of the bovine collagen X cDNA clone [10] shows limited sequence similarity with the upstream promoter sequence of the mouse gene (Figure 2). No promoter sequences are available for collagen X genes in other species. The G+C content of the mouse promoter region is unusually low: 49% between the two transcription start sites, and only 28% within the first 220 bp upstream of the 5' transcription start site. This contrasts strikingly with the corresponding region of another cartilage-specific mouse gene, the collagen II gene, where the G+C content exceeds 70% [26]. This may reflect the differential regulation of these genes. Although transcripts for these two collagen types have been localized in the same chondrocytes in chick sternum [27] and human growth plate [28], it is likely that their regulation is very different, since their expression patterns in the growth plate overlap only slightly [28].

Intron sequences

The two introns of the mouse collagen X gene exhibit a higher A+T content (66 and 62% for introns 1 and 2 respectively) than the exons (45, 55 and 42% for exons 1, 2 and 3 respectively). Intron 2 contains two short repetitive sequences of the dinucleotide repeat type. One consists of sequence (ca)₂₅(ga)₂₅, the other of (ca)₂₀. The lack of intron sequence data from other species prevented comparative analyses on conservation of intron sequences. The sequences at the intron-exon boundaries of the mouse collagen X gene conform well with the general splice consensus sequences: exon1/gtaagcc...ctttcag/exon2/gtaagct...tcttacag/exon3.

Coding sequences

The mouse nucleotide sequence is 85% identical to the human, 82% identical to the bovine and 68% identical to the chick coding sequences. The differences occur most frequently in the third position, and the overall amino acid identities between the corresponding proteins are somewhat higher: 88% between mouse and human, 82% between mouse and bovine, and 71% between mouse and chick $\alpha 1(X)$ collagen chains (Figure 4). The conservation of the nucleotide sequence is the highest in the triple-helical (COL) domain (sequence similarities between mouse and human, bovine and chick are 85, 82 and 70% respectively) and in the C-terminal NC1 domain, where the corresponding values are 87, 83 and 67% respectively. A lower degree of similarity is seen in the N-terminal NC2 domain, where the similarities between mouse and human, bovine and chick are 84, 80 and 53% respectively. A similar pattern of identities is also seen in many fibrillar collagen genes [7]. Interestingly the 3'-untranslated sequence of collagen X mRNA from the translation stop codon to the polyadenylation signal also exhibited considerable identity between the mammalian species (60–63%), whereas the chick 3'-untranslated sequence did not. The nucleotide sequences between mouse, human and bovine transcripts are particularly well conserved (73–78% identity) within the 160 nucleotides preceding the polyadenylation site located between nucleotides 8088 and 8093 (Figure 2). The other

1	tegatactaa	aacacttaag	gatatattctg	taaggtctgtg	aaagagaaaa	ccaactactt	acacggatgg	agaccatgtt	tatttctttg	ggagaaaagc
101	ctaattggga	cgtctcgaga	tcctcatagg	aaattgcacc	agtagtcaac	tggattttta	aaaggcaaa	cttgaggatt	tttttttccc	tttgaattga
201	atgtagcaaa	cttatgtaag	cacggaatag	gattattagt	taacagctct	ttcaattata	tgggaaaaatg	aaaactaggg	gagcgtctaa	ggccactttg
301	tgacctttgt	gcagctgtta	agtaaagaaa	glaaaccttc	cagggatact	gaacagccaa	ctgtcataag	tcacgggtgt	cttgcaactg	ctgtgacaag
401	tttaaaatat	ttaatatgac	tatacctgaa	atatttaatg	ctatcttttt	catgcaccag	cttctaagag	ctttccctaa	aatcctgata	tgcaaaagaa
501	tataccaata	ttttccccc	tgccccctgg	gctttgtctc	caagtttagca	aacacttagg	taagcgattt	ttacagaact	tttttcccta	ataactgaag
601	gactaacatg	atgattttaga	tctatatctc	ccccaaaagg	cgtctcataat	ttttgtatat	taccaaatat	tttcagtcac	ataacacaag	aatgtatttt
701	aaaaataaaa	agggtgaatc	atcattccat	catgaaccaa	catltagagt	agaactccta	naaggaaaa	agaaaaaaa	aaaaaatcat	gcacagccga
801	agctattaat	<u>atataatgga</u>	gacaaaaggt	ttatttttca	atgagataaa	caaggaaaaa	AGCCTGATTT	TGTACGCTG	CCCGTTAGGA	CTTCCCAACA
							C.GA....AC	.AA.A....	.AA..G.CA.	-----
901	TAATTAGTGC	TTCTTGCCCC	TGAGAGGAGG	AGCTTCGGCT	CAGGGGAATC	TCATGCAATA	AGGGAAGAAA	ACAGTATAAA	TACTCCAGGG	CAGAAGTGGG
							G	G	GG.CTT	..T...AA.TT
1001	GAAGGCATTA	TCCACTGCCTC	CTGGGCAGAG	-GAA--GCCA	GGAAAGCTGC	CCCACGCATC	TCCCAG-CAC	CAGgtaagcc	tcgtctgagg	gtcaatgggc
	A..AA.T..	..GAG..T..A	AA.....CTA	TC..CT.GT.	A....CCAC.	T....G...	...G...AGT.	.G.		
	INTRON 1 (563 bp)									
1601	gctatgtgta	atgattctct	tctgtctttc	agAATCTATC	TGAAAAATAG	CTGCCTCAAA	TACCCTTTCT	GCTGCTAATG	TTCTTGACCC	TGGTTCATGG
				C.G.	...G.C....	A....	C.G.AC..T.		.C....A.T	
1701	CATGTTTAT	GCTGAACGGT	ACCAACGCC	CACAGGCATA	AAGGGCCAC	TTGCTAGCCC	CAAGACACAA	TACTTCATCC	CATACGCCAT	AAAGAGTAAA
	A3.....	A....G..A.	A..	T.....	..A.....	CAT.C.C.A.	C.G	.T.....	.G..T....	...G....
1801	Ggtaagctca	actgtacatt	ttgtctatt	tcccagacag	INTRON 2 (3414 bp)					
5201	tggtgtgtct	tacagGGATT	CCAGTAAGAG	GAGAACAAGG	CATTCTGGT	CCACCAGGCC	CAACCGGACC	TCGAGGACAC	CCAGGTCCCT	CAGGACCGCC
		.TG.A.T..	C.G....	G....	...C....		.CG.T....	G...	C..G.	A..
5301	AGGCAAGCCA	GGCTATGGAA	GTCTGGACT	CCAAGGAGAG	CCAGGGTTGC	CAGGACCACC	AGGAATATCA	GCCACGGGGA	AGCCAGGCC	GCCAGGCCCG
	A....	..ACC....	..G.C....	C....			G....CCG...	T....	TT....	A.TC
5401	CCAGGCAAA	CAGGGGAGAG	AGGACCATAT	GGACACAAAG	GAGATATTGG	CCCAGCTGGC	TTACCCGGAC	CTCGGGGCC	TCCAGGGGCC	CCTGGAATTC
	A....	A.....	..T.A....	CA....		A.....T	A.....A....	A.....A....	A.....A....	..C.....
5501	CTGGCCAGC	TGGAATTTCT	GTGCCAGGAA	AACCTGGACA	GCAGGGACTT	ACAGGTGCC	CAGGACCTAG	GGGCTTTCT	GGAGAGAAGG	GTGCACAAGG
	..C.....		G.C.	A..A.....	AG.A.		T..C.			
5601	AGCCCTGGT	GTGAATGGGC	GGAAAGGGGA	AACAGGATAT	GGCTCTCTCT	GGCTTCCAGG	TGAGAGGGGT	CTTCCAGGCC	CTCAAGGTCC	CATAGGACCC
	G.T.....	C.....A.	A.....	..TG..GC..	T.TA.....			..C.....T.	G.....	..CG....A
5701	CTTGGCCCT	CTGGAGTGGG	AAGAAGAGGT	GAAACGGCT	TTCCAGGACA	GGCCGGCATC	AAAGGTGACC	GGGGTTTCCC	AGGACAATG	GGACCATCAG
	TC		G.A.....	T.G.	A.....	..A....C..	T.A.	A.....G.T..	G....G.	...G..G.C.
5801	GTCCACCAAG	TCCCAAGGT	CCTCCCGGA	AGCAAGGACG	AGAAGGTATT	GGGAAGCCAG	GAGCCATTGG	ATCCCTTGGT	CAGCCAGGTA	TTCCAGGAGA
	..C..T....	..C..T....	..C..T....	..C..T....	..C..T....	..C..T....	..C..T....	..C..T....	..C..T....	..C..T....
5901	AAAAGGCCAC	CCAGGGGCTC	CAGGAATAGC	TGGTCTCTCA	GGAGCTCTCT	GCTTTGGAAA	ACAAGGCTTG	CCAGGTTTGA	GGGGACAAAG	GGGACCTGCT
	G.....T..G	C....	G.TC..G	G....	G....	G....	C....	C....	C....	TG
6001	GGTCTTCTC	GGGCTCCAGG	TGCCAAAGGG	GAACGAGGGC	CACAGGCTCA	TCCTGGAGAA	CCGGGTCTGC	CTGGATCCCC	TGGGAATATG	GGACCCCAAG
	..C.....A.	..AG....	A....	C....	G....	G.A.....	C....	C..T.	A.....	
6101	GACCTAAAGG	AATCCAGGG	AACCATGGCA	TTCCAGGCGC	TAAAGGTGAG	ATAGGTCTAG	TTGGGCTTGC	AGGCCCCCT	GGGGCTAGAG	GAGCAAGGGG
	..C..A....	..C..A....	..C..A....	..C..A....	..C..A....	..C..A....	..C..A....	..C..A....	..C..A....	..C..A....
6201	TCCACCTGGG	TTAGATGGAA	AAACAGGGTA	TCCTGGGGAG	CCAGGTCTCA	ATGGTCTTAA	GGGTAAACCA	GGGTATCCAG	GACAAAAGG	TGATCTTGGA
	CT..CT....	C....	..C..A..A.	C....	C....	G....	G....	G....	C....	C....
6301	CTGGGAGGAA	CCCTGGGCT	TCGAGGTCT	GTGGGCTCT	TAGGAGCTAA	AGGAGTCCCT	GGACACAATG	GTGAGGAGG	TCCAAAGGTT	GAACCTGGAA
	A.T.C....	..T....	..C..C..C.	..G..T..A.				..C..A..T.	..G..C....	..TC....A.
6401	TACCAAGTAC	CAGGGGCCCT	ACTGGGCCAC	CAGGTGTCTC	AGGATTCCTC	GGATCTAAGG	GTGACCTTGG	AAACCCAGGT	GCTCCAGGCC	CAGCTGGCAT
	T..A.T.	..T....	..C..A..T.	C..T.	C..T.	C..T.	C..T.	C..T.	C..T.	C..T.
6501	AGCAACTAAG	GGCCTCAATG	GGCCCACTGG	TCCTCCAGGC	CCTCTCTGGT	CAAGAGGCCA	CAGTGGAGAA	CCTGGTCTCC	CAGGTCTTCC	GGGTCCCCCA
	GTC..	T....	..A..T....	..G..A....	G..G.	A....	TGC....G..G	C....	G..C..	A..G....
6601	GGACCCCCCG	GCCAAGCAGT	CATGCCCTGAT	GGCTTCATAA	AGGCAGGCCA	GAGGCCCCAGG	CTTCTTGGGA	TGCCGCTTGT	CAGTGCTAAC	CACGGGGTAA
	..C..T..A.	T..C	..C..C..G	..A..TG..		A.....GTTT		T.....C..T	G..A....	
6701	CAGGTATGCC	CGTGTCTGCT	TTTACTGTCA	TTCTCTCTAA	AGCTTACCCA	GCAGTAGGTG	CCCCCATCCC	ATTGTATGAG	ATTCTGTACA	ATAGGCAGCA
	A....	..T....	..C....	..C....	..TA....	A..T..T.		..C..A..	..T..A..T.	..C..A..A.
6801	GCATTACGAC	CCAAGATCTG	GTATCTTTAC	CTGTAAGATC	CCAGGCATAT	ACTATTTCTC	CTACCACTG	CATGTGAAAG	GGACTCAGT	TTGGGTAGGC
	T....	A....	A....	G....	G....	..T..C....	..T..A..A.	C..T.	C..T.	T
6901	CTGTATAAGA	ACGGCACGCC	TACGATGTAC	ACGTATGATG	AGTACAGCAA	AGGTACCTG	GATCAGGCTT	CAGGGAGTGC	AATCATGGAG	CTCACAGAAA
	C....	..GTA....	A....	T....	T....	G....	G....	..C..G..C..	C..G.	C..G.
7001	ATGACCAAGT	ATGGCTCCAA	TTGCCCAATG	CAGAATCAAA	CGGCCCTTAC	TCCTCTGAGT	ACGTCCACTC	GTCTCTCTCA	GGATTCCTAG	TGGCTCCCAT
	..C.....G	G	GC..G.	G..G..T	C....	C....	C....	C..T.	T..G.	A..
7101	GTGAGCACTT	CCTGTCAAGC	TCATCTTATT	CTCCGCTTAG	AAAGGCTTTC	CACCCAATTC	CATCGCGCCC	TCCTTGAGGA	TGCATTTGGA	GGTAGGCTAA
										
7201	AAGTAATGTG	AACAACCTAC	ACTTCTTAAA	GTGCACTCTG	GSACCAAGTT	TCCTCACTTC	TAAAGTGAAC	AGTAACAGTG	GGTATGTGAG	GATCCTCCCG
7301	GGAGCACTCA	GCAATCCTAA	AGCCCTGGGG	GGTGGTGAGG	GGGTGGTGTC	TGTTATTCCA	TCAATACTTA	CATTTTTPPA	ACAGAAGCTT	TAAAAAAG
7401	CCTCACAATA	ACAGAAATAA	GACTCTAACT	CACCATAAAT	CTCATTTCAA	AAACTCAACA	TTCTCCTTTC	AAAAAAGTCC	ATTTTAAATA	ATATGAAAAC
7501	ATCPAGGAAA	CATGCAGGAA	TTATCATGCA	ACTTTATATT	CTTTAAATAC	TTGAATATCA	AATTGAAATG	ACACTGTATA	CCTAGGATAT	TTCTGATGGT
7601	GCATCACTCA	ATGGCTCATA	CAACCCCTTT	CATCAAGATA	AAATCAAAATG	CACATGGTAC	CACATGGTTC	TTGAAGTGGC	CATAAGAAAC	CAAAATACGT
7701	TATTAAATAT	AATTTCTGAT	ATAAGGTGCT	TTGTACAGC	TGTTTCAAGC	TTTTCTGCAA	TTGCAGAAAG	TCCACATACC	TGGTGAATCT	TGAAACAAG
7801	TGCTGACTT	ATGTCAGTAA	CACAGTGCAC	AGTGTGATTA	CTCTGATTTA	ATTCCTTGCA	TCTTTGTAAA	ATGATTTTAT	GGGCTTATAC	AATATTAGAA
7901	TAGACACTTT	TGTGCTCTTT	CAATCGAGTG	GAAGGATATA	CATTAAAGGT	TTGAACATCA	GACTAGAAAC	GAAAGACAT	TATTTATTTA	TGCTCTGTAC
8001	TGTATTTTAA	CATTATTGTT	TAAAACTGGT	AGGCTGTACC	TCATTATTAA	AAGCCCGGAG	TATTTAATCT	ACTTGGTATT	ACGAAGCAAT	AAATGACAT
8101	GAATAGACCT	TTATTCTGTG	CTGAGTCTTT	TCATAGTGAC	AACAGGCTGT	TCACAGCTTT	ACAGCTTCCG	ACTGGCTGAA	GCTAATAAGA	AAAGCAGTGA

Figure 2 Nucleotide sequence of the mouse collagen X gene

The sequence shown covers 0.86 kb of the promoter, all three exons (upper case letters), exon-intron boundaries and 1.1 kb of the 3'-untranslated region. The top line of each pair is the nucleotide sequence of the mouse gene, and the bottom line is the bovine exon sequences when different from the mouse sequence. Nucleotide identity is shown with a full-stop (.), and gaps with dashes (-). Underlining (in the promoter) highlights the CCAAT box and the two TATAA boxes, the translation start site (in exon 2), the translation stop codon and the two polyadenylation signals (in exon 3). In exon 1 the nucleotides found only in the longer $\alpha 1(X)$ collagen transcript are shown in italics. The boxed areas denote NC2 and NC1 domains. Numbering of the mouse nucleotides on the left includes also intron sequences.

Table 1 Comparative analysis of the chick, bovine, human and mouse collagen X genes

The sizes of all domains are given in base pairs. The size of the 3'-untranslated (3'-UT) sequence is given from the translation stop codon to the end of the polyadenylation signal (AATAAA). Data for the chick gene is from refs. [9] and [25], for the bovine gene from ref. [10], and for human gene from ref. [11]. *The exon size estimates for the bovine gene are based on the assumption that the exon-intron boundaries correspond to those in the mouse, human and chick genes.

Gene...	Size (bp)			
	Chick	Bovine	Human	Mouse
Exon 1	97	141*	NA	61/207
Intron 1	670	NA	NA	563
Exon 2	159	169*	169	169
Intron 2	~ 2000	NA	~ 3200	3414
Exon 3 (coding)	1889	1868*	1886	1886
Exon 3 (3'-UT)	240	955	1041	993
Exon 3 (total)	2136	2824	2927	2914

Figure 3 Determination of the transcription start site

(a) For primer extension of $\alpha 1(X)$ collagen mRNA, two 30-mer oligonucleotides within exon 1 were synthesized and used to prime cDNA synthesis by reverse transcriptase. (b) The reaction product obtained with primer B was resolved on a denaturing sequencing gel with the corresponding sequencing reaction run on the side. (c) A corresponding electrophoretic fractionation of the reaction products and sequencing reactions obtained with primer C. The locations of the TATAA sequences are shown in both panels.

AATAAA sequence in the mouse located at nucleotides 7935–7940 is not conserved in the bovine and human sequences.

N-Terminal non-collagenous domain (NC2)

In the mouse, this domain consists of 56 amino acids (Figure 4). The size is the same in bovine and human collagen X [10,11],

NC2 domain:

```

1 MLEPQIPFLLE MFLTLVHG-- MFYAERYQTF TGKQPLASPT KQYFIFPFI KSKSIPVR
  ...S.N... V...T... PNT ...F...T... A...A...
  ...TAL... S.N... V...T... PSNT ...F...T... G...VSL...
  ...SL... PC.N... SD GYFS...KQ SS...F... H...F...V... Q...VGM...

```

Triple helical domain:

```

1 GEQGIQPGPG PTGPRSHGPG SGPGRGPGY SPGLQGEPL PGPGISATG KPGGLPGPGK
  ...T... A...T... ...P...V... V...L...
  ...P... A...T... ...P...P... ...P...L...
  ...P... I...Q... A...T... P...P... ...P...TV... R...R...

```

```

61 PGRGPGYSHK GEIGPAGLPG PRGPGPGGI PGFAGISVPG KPGQQLTGA PGPRGPGGK
  ...P... V...T... ...P...P... ...P...P...
  ...LN...P... ...P...P... ...P...P...
  ...LN...E... EA...V... A...Q... ...L...P... PP... Q...R...

```

```

121 GAQAPGVNG KEGTGYGSE GRIGERGLEG PCHIGPGPG SGVGRGNG FPGPGGIGD
  ...P...M... Q...M...A... ...T...S... P...K... V...L...
  ...T...L... Q...M...H... C...T...T... P...K... L...L...
  ...P...I... Q...M...F...V... ...N... ...L...P... A...I...K... L...M...

```

```

181 PGRPGMGPS GPGPGPGPG KGRGPGGPG GAGPGPGPG IPGKPGGPA PGAGPGGAP
  ...P...I... ...P...P... ...A...A... ...T...L... ...P...P...
  ...V...R...A... ...S...P... ...P...L... ...M...Q... ...T...L...
  ...L...A...E... ...P...P... ...P...P... ...A...L...P... A...P...S...L...

```

```

241 GPGQGLDGL GGRPGAGLPG GAGPGGPGPG PGRGPGGPG PGPGGPGPG GPGPGGPGG
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...

```

```

301 PGRGPGGPG GPGPGPGPG GAGPGGPGPG KPGPGGPGPG KPGPGGPGPG KPGPGGPGPG
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...

```

```

361 PGRGPGGPG GPGPGPGPG GAGPGGPGPG KPGPGGPGPG KPGPGGPGPG KPGPGGPGPG
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...

```

```

421 PGRGPGGPG GPGPGPGPG GAGPGGPGPG KPGPGGPGPG KPGPGGPGPG KPGPGGPGPG
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...
  ...P...P... ...P...P... ...P...P... ...P...P... ...P...P...

```

NC1 domain:

```

1 VMPPGFIKAG QRPLCGMPL --VGNHGVV GMVSAFTVI LSKAYFAVGA PIPDEILYN
  ...E...T... ...S...Q... ...P...T... ...I...T... ...K...
  ...L...V... ...P...T... ...P...K... ...P...T... ...K...
  ...T...V...G...S...E...S...F...M...K...Q...L... ...G...T...V...K...K...

```

```

61 RQGHYDFRSG IPTCKIFGIY YFSYHVHVKG THVWGLYEN GTPTMTYDE YSKGYLQAS
  ...T...Q... ...T...I... ...A...A... ...S...V... ...Q...
  ...T...T... ...P...L... ...P...A... ...N...A... ...S...V... ...Q...
  ...T...T... ...P...L... ...P...A... ...N...A... ...S...V... ...Q...

```

```

121 GSAIMELTEN DQWLQLPNA KNGLYGSEY VHSFSGFLV APY
  ...I... ...G... ...P... ...F...
  ...V... ...I... ...G... ...P...
  ...M... ...S... ...G... ...F...

```

Figure 4 Deduced amino acid sequence of the mouse $\alpha 1(X)$ collagen

The top line shows the amino acid sequence of the different domains of the mouse protein. The corresponding human (second line), bovine (third line) and chick (bottom line) amino acid sequences are also shown when they differ from the mouse sequence [9–11, 25]. The arrow marks the putative signal peptidase cleavage site. The conserved discontinuities of the triple-helical domain are marked with a line above the mouse sequence, and the 13 conserved tyrosine residues of the NC1 domain are indicated with asterisks.

while in the chick protein this domain contains 52 amino acids. The cleavage site of the signal peptidase has been determined in the chick [9]. By sequence comparison it appears that all species have a signal peptide of 18 amino acids and an NC2 domain of 38 (in chick 34) amino acids. As with fibrillar collagens the divergence of the collagen X molecule is the greatest in the N-terminal non-collagenous domain: the amino acid identity between mouse and human is 84%, between mouse and bovine 71%, and between mouse and chick only 41%.

Collagenous domain (COL)

The size of the single triple-helical domain of the mouse collagen X gene is 463 amino acids. The size is the same in the bovine and human proteins, but three amino acids shorter in the chick [8–11]. At the amino acid level the identity between mouse and human COL domains is 87%, that between mouse and bovine

Figure 5 Localization of $\alpha 1(X)$ collagen mRNA in the epiphyseal growth plate by *in situ* hybridization

Formalin-fixed and paraffin-embedded samples of newborn mouse limbs were sectioned and pretreated for *in situ* hybridization with probes for mouse $\alpha 1(X)$ collagen mRNA (1.2 kb *Bam*HI fragment from exon 3) (panels **a** and **c**), and for the $\text{pro}\alpha 1(\text{II})$ collagen mRNA (insert from clone pMCol2a1-1) (panels **b** and **d**). After hybridization the sections were washed, the probe detected by autoradiography, and the sections stained with haematoxylin and eosin. Panels **(a)** and **(b)** represent phase-contrast micrographs of the epiphyseal growth plate areas, and panels **(c)** and **(d)** are the respective dark field images. The epiphysis and articular surface are to the left and the growth plate and osseochondral junction to the right. Collagen X mRNA is located in hypertrophic chondrocytes of the growth plate and the secondary ossification centre of the epiphysis. The bar in **(a)** corresponds to 100 μm .

Figure 6 Localization of $\alpha 1(X)$ collagen mRNA in the vertebral column by *in situ* hybridization

Formalin-fixed and paraffin-embedded sections of newborn mice were hybridized *in situ* with probes for mouse $\alpha 1(X)$ collagen mRNA (panels **a** and **c**), and for the $\text{pro}\alpha 1(\text{II})$ collagen mRNA (panels **b** and **d**), as described in Figure 5. Panels **(a)** and **(b)** are phase-contrast micrographs of lumbar vertebral column with the cartilaginous spinous process to the left and the neural arch to the right, with an ossification centre in between. Panels **(c)** and **(d)** show the respective dark-field images. Collagen X mRNA is located in hypertrophic chondrocytes in the cartilaginous growth zone of the spinous process and in the central calcification centre of the arch. The bar in **(a)** corresponds to 100 μm .

83 %, and that between mouse and chick 73 %. The locations of the discontinuities in the triple-helical structure are also conserved between the species: the repeating GXY triplet has been replaced by GX in five locations and by GXYXY in three cases (Figure 4). In the bovine protein, one of the GX type discontinuities has been replaced by a GHCTPCRP sequence [10], which provides the cysteine residues for the interchain disulphide bonds seen in bovine collagen X molecules, but not in those of the chick, human or mouse [2].

Two cleavage sites for interstitial collagenases have been identified in the chick collagen X molecule [15]. In both cases

cleavage occurs in longer discontinuity domains of the GXYXY type at Gly⁹²-Ile⁹³ and Gly⁴²⁰-Leu⁴²¹ bonds. Both cleavage sites are conserved in the mouse, bovine and human genes as Gly⁹⁵-Ile⁹⁶ and Gly⁴²³-Ile⁴²⁴ bonds. In the mouse gene the third long discontinuity (amino acids 45–49) also codes for the recognition sequence (GISAT). This cleavage site is not, however, conserved in the chick, bovine or human genes.

C-Terminal non-collagenous domain (NC1)

The size of this domain in the mouse and human $\alpha 1(X)$ collagen chains is 161 amino acids, in bovine and chick it contains 155 and 162 amino acids respectively. Apart from the first 24 amino acids, this domain exhibits a high degree of identity between mouse and human (93 %), bovine (88 %) and chick (81 %) collagen X chains. The overall identity of the NC1 domain between mouse and human is 91 %, between mouse and bovine 83 %, and between mouse and chick 76 %. This conservation would favour the suggestion that the association of the $\alpha 1(X)$ chains begins at the NC1 domain. Typically the locations of all 13 tyrosine residues, as well as of the single putative N-linked carbohydrate attachment site (amino acids 99–101) and the single cysteine residue (amino acid 74), are conserved between the four species studied (Figure 4).

Expression of the collagen X gene

Depending on the start site of transcription the sizes of the mRNAs coding for the mouse $\alpha 1(X)$ collagen are expected to be between 3.2 and 3.3 kb plus the poly(A) tail. In Northern analysis the probes used recognized transcripts migrating at approx. 3.3 kb present only in total RNA extracted from cartilage (results not shown; see [19]). *In situ* hybridization was employed to further analyse the cell type responsible for collagen X production in newborn mouse fetuses (Figures 5 and 6). The results clearly show that the expression of the $\alpha 1(X)$ collagen mRNA is restricted to hypertrophic chondrocytes located at growth plates and ossification centres of epiphyses and vertebrae.

After this paper was submitted, Apte et al. [29] published a partial gene sequence for mouse collagen X. Their sequence contains 13 nucleotide differences compared with ours which result in nine amino acid changes. Four of these are clustered in the NC1 domain between amino acids 50 and 55 (Figure 4), resulting from a different reading frame between nucleotides 6759 and 6775. Our sequence in this area is nearly identical with the human and bovine sequences and does not introduce an additional tyrosine residue.

The results of the present study clearly demonstrate that the structure and expression pattern of the mouse collagen X gene shares similar unique features seen in the chick, bovine and human genes. This information suggests that the functions of collagen X in the mouse are analogous to those in other species, and can now be studied by generating specific mutations in the gene and expressing them in transgenic mice. Our experience with the generation of transgenic mice harbouring mutations in the homotrimeric collagen II gene has confirmed the usefulness of this approach [30]. With the completion of the genomic structure of the mouse collagen X, such experiments have now become possible also with this collagen type.

This study was financially supported by the Medical Research Council of the Finnish Academy and by National Institutes of Health Grant AR 40335. The expert technical help of Merja Lakkisto, Tuula Oivanen and Aila Rusi is gratefully acknowledged.

REFERENCES

- 1 Schmid, T. M. and Linsenmayer, T. F. (1985) *J. Cell Biol.* **100**, 598–605
- 2 Schmid, T. M. and Linsenmayer, T. F. (1987) in *Structure and Function of Collagen Types* (Maynes, R. and Burgeson, R. E., eds.), pp. 223–259, Academic Press, New York
- 3 Stockwell, R. A. (1979) *Biology of Cartilage Cells*, Cambridge University Press, Cambridge
- 4 Hunziker, E. B., Schenk, R. K. and Cruz-Orive, L.-M. (1987) *J. Bone Jt. Surg. Am. Vol.* **69-A**, 162–173
- 5 Castagnola, P., Moro, G., Discalzi-Cancedda, F. and Cancedda, R. (1986) *J. Cell Biol.* **102**, 2310–2317
- 6 Adams, S. L., Pallante, K. M., Niu, Z., Leboy, P. S., Golden, E. B. and Pacifici, M. (1991) *Exp. Cell Res.* **193**, 190–197
- 7 Vuorio, E. and de Crombrughe, B. (1990) *Annu. Rev. Biochem.* **59**, 837–872
- 8 Ninomiya, Y., Gordon, M., van der Rest, M., Schmid, T., Linsenmayer, T. and Olsen, B. R. (1986) *J. Biol. Chem.* **261**, 5041–5050
- 9 LuValle, P., Ninomiya, T., Rosenblum, N. and Olsen, B. R. (1988) *J. Biol. Chem.* **263**, 18378–18385
- 10 Thomas, J. T., Kwan, A. P. L., Grant, M. E. and Boot-Handford, R. P. (1991) *Biochem. J.* **273**, 141–148
- 11 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
- 12 Yamaguchi, N., Mayne, R. and Ninomiya, Y. (1991) *J. Biol. Chem.* **266**, 4508–4513
- 13 Muragaki, Y., Jacenko, O., Apte, S., Mattei, M.-G., Ninomiya, Y. and Olsen, B. R. (1991) *J. Biol. Chem.* **266**, 7721–7727
- 14 van der Rest, M. and Garrone, R. (1991) *FASEB J.* **5**, 2814–2823
- 15 Welgus, H. G., Fliszar, C. J., Seltzer, J. L., Schmid, T. M. and Jeffrey, J. J. (1990) *J. Biol. Chem.* **265**, 13521–13527
- 16 Schmid, T. M. and Linsenmayer, T. F. (1990) *Dev. Biol.* **138**, 53–62
- 17 Kwan, A. P. L., Cummings, C. E., Chapman, J. A. and Grant, M. E. (1991) *J. Cell Biol.* **114**, 597–604
- 18 Chen, Q., Gibney, E., Fitch, J. M., Linsenmayer, C., Schmid, T. M. and Linsenmayer, T. F. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8046–8050
- 19 Elima, K., Metsäranta, M., Kallio, J., Perälä, M., Eerola, I., Garofalo, S., de Crombrughe, B. and Vuorio, E. (1992) *Biochim. Biophys. Acta* **1130**, 78–80
- 20 Devereux, J., Haerberli, P. and Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395
- 21 Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. and Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299
- 22 Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn., pp. 7.58–7.83, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
- 23 Sandberg, M. and Vuorio, E. (1987) *J. Cell Biol.* **104**, 1077–1084
- 24 Metsäranta, M., Toman, D., de Crombrughe, B. and Vuorio, E. (1991) *Biochim. Biophys. Acta* **1089**, 241–243
- 25 Ninomiya, Y., Castagnola, P., Gerecke, D., Gordon, M. K., Jacenko, O., LuValle, P., McCarthy, M., Muragaki, Y., Nishimura, I., Oh, S., Rosenblum, N., Sato, N., Sugrue, S., Taylor, R., Vasios, G., Yamaguchi, N. and Olsen, B. R. (1990) in *Extracellular Matrix Genes* (Sandell, L. J. and Boyd, C. D., eds.), pp. 79–114, Academic Press, San Diego
- 26 Metsäranta, M., Toman, D., de Crombrughe, B. and Vuorio, E. (1991) *J. Biol. Chem.* **266**, 16862–16869
- 27 LuValle, P., Hayashi, M. and Olsen, B. R. (1989) *Dev. Biol.* **133**, 613–616
- 28 Reichenberger, E., Aigner, T., von der Mark, K., Stöß, H. and Bertling, W. (1991) *Dev. Biol.* **148**, 562–572
- 29 Apte, S. S., Seldin, M. F., Hayashi, M. and Olsen, B. R. (1992) *Eur. J. Biochem.* **206**, 217–224
- 30 Garofalo, S., Vuorio, E., Metsäranta, M., Rosati, R., Toman, D., Vaughan, J., Lozano, G., Mayne, R., Ellard, J., Horton, W. and de Crombrughe, B. (1991) *Proc. Natl. Acad. Sci. U.S.A.* **88**, 9648–9652