Structure of cDNA clones coding for the entire prepro α 1(III) chain of human type III procollagen

Differences in protein structure from type I procollagen and conservation of codon preferences

Leena ALA-KOKKO, Sirpa KONTUSAARI, Clinton T. BALDWIN, Helena KUIVANIEMI and Darwin J. PROCKOP*

Department of Biochemistry and Molecular Biology, Jefferson Institute of Molecular Medicine, Jefferson Medical College, Thomas Jefferson University, Philadelphia, PA 19107, U.S.A.

Two overlapping cDNA clones that cover the complete length of the mRNA for human type III procollagen were characterized. The data provided about 2500 base pairs of sequence not previously defined for human type III procollagen. Two tripeptide sequences of -Gly-Xaa-Yaa- were identified that were not detected previously by amino acid sequencing of human type III collagen. The two additional tripeptide units, together with three previously detected, establish that the α 1(III) chain is 15 amino acids longer than either the α 1(I) or α 2(I) chains of type I collagen. The additional tripeptide units made hydrophathy plots of the *N*-terminal and *C*-terminal regions of type III collagen distinctly different from those of type I collagen. The data also demonstrated that human type III procollagen has the same third base preference in codons for glycine, proline and alanine that was previously found with human and chick type I procollagen. In addition, comparison of two cDNA clones from the same individual revealed a variation in structure in that the codon for amino acid 880 of the α 1(III) chain was -CTT- for leucine in one clone and -TTT- for phenylalanine in the other.

INTRODUCTION

Type III collagen is a member of the family of fibrillar collagens and accounts for 5–20% of the collagen present in mammals (see Mayne & Burgeson, 1987; Miller, 1988). It is generally found in association with type I collagen, but is particularly abundant in tissues such as aorta. In contrast, it is absent from bone. The function of type III collagen has not been fully defined, but its essential role is illustrated by the heritable disorder of connective tissue known as Ehlers–Danlos syndrome type IV in which decreased synthesis or synthesis of structurally abnormal type III procollagen produces aortic aneurysms and rupture of hollow organs (see Prockop & Kivirikko, 1984; Byers & Bonadio, 1985). The structure of the α 1(III) chain of type III collagen was first defined by partial amino acid sequencing of bovine type III collagen (Glanville & Fietzek, 1976; Fietzek *et al.*, 1979; Brandt *et al.*, 1984) and then by amino acid sequencing of human type III collagen (Seyer *et al.*, 1980; Seyer & Kang, 1981). cDNAs covering part of the structure of chick type III procollagen have been isolated and sequenced (Yamada *et al.*, 1983*b*). Subsequently, partial cDNAs for human (Loidl *et al.*, 1984; Chu *et al.*, 1985; Miskulin *et al.*, 1986; Mankoo & Dalglish, 1988; Toman *et al.*, 1988) and for murine (Wood *et al.*, 1987) type III procollagen have been isolated and sequenced. In addition, the structure of the 5' and 3' ends of clones covering the chick type III procollagen gene (Yamada *et al.*, 1983*a,c*, 1984), the 5' end of clones for murine type III procollagen gene (Liau *et al.*, 1985), and the 3' end of clones for human type III procollagen gene (Chu *et al.*, 1985) have been reported.

In the present paper we describe two overlapping cDNAs coding for the full length of the prepro α chain of human type III procollagen. The data provide about 2500 bp of sequences not previously defined. Because we have recently reported the structure of full-length cDNAs for the prepro α 1(I) (Tromp *et al.*, 1988) and prepro α 2(I) (Kuivaniemi *et al.*, 1988) chains of human type I procollagen, the present data made it possible to carry out a detailed comparison of structural differences between these human fibrillar procollagens.

MATERIALS AND METHODS

Synthesis and screening of the cDNA library

Total RNA was isolated from one 175 cm² flask of cultured skin fibroblasts from a proband with a lethal variant of osteogenesis imperfecta. Previous work demonstrated that osteogenesis imperfecta in the proband was produced by a single base mutation that converted Gly-907 of the α 2(I) chain of type I procollagen to aspartate (Baldwin *et al.*, 1989). The RNA was isolated using the guanidine isothiocyanate/CsCl density gradient centrifugation method (Maniatis *et al.*, 1982). Approx. 15 μ g of total RNA was used to synthesize cDNA by priming the first strand with oligo(dT) by the procedure of Gubler & Hoffman (1983) using Maloney murine leukaemia virus reverse transcriptase (Bethesda Research Laboratories). *Eco*RI linkers were attached to the double-stranded cDNA, ligated to a λ vector (λ ZAP; Stratagene), and packaged *in vitro* using a commercial packaging extract (Gigapack; Stratagene). The bacteriophage library contained approx. 10⁷ independent clones. The

* To whom correspondence and reprint requests should be addressed.
The sequence data have been submitted to the EMBL/GenBank libraries.

library was screened with a 0.8 kb *PstI/PstI* fragment isolated from Hf-934, a cDNA clone coding for human type III procollagen (Chu *et al.*, 1985). The probe was labelled by nick translation to a specific activity of about 10^8 c.p.m./ μ g of DNA and the library was screened by plaque hybridization.

Thirty positive clones for the gene were plaque-purified and the inserts in the vector were automatically converted to a phagemid with a helper virus (R408) under conditions recommended by the supplier (Stratagene). Deletion libraries were prepared from the clones using exonuclease III and S1 nuclease with a commercial kit (Promega). Double-stranded DNA was sequenced by the dideoxy method (Sanger *et al.*, 1977; Messing, 1983) using T7 DNA polymerase (Sequenase; United States Biochemical Corp.).

RESULTS AND DISCUSSION

Restriction map and nucleotide sequence of the cDNAs

One of the cDNA clones (S413) isolated from the human skin fibroblast library was essentially a full length copy of the mRNA (Fig. 1). The 5' end of the cDNA began in the 5'-non-translated region of the message and the 3' end was at the *EcoRI* site that is found near the 3'-end of the coding sequences (Loidl *et al.*, 1984). A second large cDNA (S31) that overlapped the first and extended to the 3'-end of the mRNA was obtained from a second cDNA library that was prepared with methylated double-stranded cDNA. The two cDNAs essentially covered the full length of the message (Fig. 1). Most of the nucleotide sequencing was carried out on these two large cDNAs. In addition, several smaller cDNA clones were also partially sequenced. With the one exception noted below, all overlapping regions that were examined had the same

nucleotide sequence. Also, except as noted below, the nucleotide sequences were identical to previously published nucleotide sequences in instances where there was overlap and the nucleotides were examined. The identity of the clones was further identified by Northern blot analysis in which one of the large clones (S413) was shown to hybridize the two mRNA bands about the same size as those previously detected with shorter cDNAs for type III procollagen (Loidl *et al.*, 1984).

DNA sequencing of the cDNAs was carried out as indicated in Fig. 1. All regions for which no sequences were previously published were sequenced on both strands to confirm the data. The sequences included over 2500 bp not previously defined for the human type III procollagen and over 1500 bp of sequences not previously defined for any species (Fig. 2).

The data confirmed previous impressions that the amino acid sequences of human type III collagen were essentially the same as those in murine and bovine. Surprisingly, however, the data revealed two additional -Gly-Xaa-Yaa- tripeptide units not previously detected in human type III collagen. One of the tripeptides was a sequence of -Gly-Leu-Ala- at the *N*-terminus of the triple helix (amino acids 168–170 in Fig. 2). It apparently went undetected because the amino acid sequence data suggested that the residue now identified as the first glycine in the triple helix was valine (Seyer *et al.*, 1980). The second additional tripeptide unit was a sequence of -Gly-Pro-Ala- (amino acids 225–227 in Fig. 2). The tripeptide was detected by partial nucleotide sequencing of a 5'-cDNA for murine type III collagen (Wood *et al.*, 1987) but not by more extensive amino acid sequencing of the bovine and human proteins (Fietzek *et al.*, 1979; Seyer *et al.*, 1980; Brandt *et al.*, 1984). Apparently for this reason, no correction was made previously for this

Fig. 1. Sequencing strategy used and partial restriction map of the cDNA clones for human type III procollagen

Top of figure indicates the *N*-propeptide (N), triple-helical domain and C-propeptide (C) of type III procollagen. Second line indicates the partial restriction map of the cDNAs. Third line indicates two overlapping cDNA clones for type III procollagen (S413 and S31). The lower part of the figure indicates the regions that were sequenced. Symbols: B, *Bam*HI; E, *Eco*RI; H, *Hind*III; P, *Pvu*II; Ps, *Pst*I; S, *Sac*I; Sa, *Sal*I; Sp, *Sph*I; X, *Xho*I; Xb, *Xba*I.

[illegible]

mouse	GGA	CCT	G	CAG	GGA	CAC	T	GCT	GGT	GCT	CAA	GGT	CCT	CCT	T	GGC	CCT	C	GGG	ATT	AAT	GGT	AGT	CCT	GGT	GGT	A	G	T	1194		
human	Gly	Pro	Gln	Gly	His	Ala	Gly	Ala	Gly	Ala	Gln	Gly	Pro	Pro	Gly	Pro	Gly	Pro	Gly	Ile	Asn	Gly	Ser	Pro	Gly	Gly	Lys	Gly	Glu	Met	398	
mouse	-	-	-	-	-	-	-	-	-	-	-	Pro	-	-	-	-	-	-	-	Asn	-	-	-	-	-	-	-	-	-	-	-	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Ser	Asp	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Ser	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGT	CCC	A	GCT	GGC	ATT	CCT	GGG	GCT	CCT	GGA	CTG	ATG	GGA	A	A	GGC	CGG	GGT	CCT	CCA	GGA	CCA	GCC	GGT	GCT	AAT	GGT	GCT	ATA	C	1275
human	Gly	Pro	Ala	Gly	Ile	Pro	Gly	Ala	Gly	Ala	Pro	Gly	Leu	Met	Gly	Ala	Arg	Gly	Pro	Pro	Gly	Pro	Ala	Gly	Ala	Asn	Gly	Ala	Pro	Gly	Met	425
mouse	-	-	-	-	-	-	-	-	-	-	-	-	Ile	-	-	-	-	-	-	-	-	-	Pro	-	-	Thr	-	-	-	Val	-	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	Ile	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGA	CTG	ACA	CGA	GGT	GGT	GCA	GGT	GAG	CCT	GGT	AAG	AAT	GGT	GCC	AAA	GGA	GAG	CCC	GGA	CCA	CGT	GGT	GAA	CGC	GGT	GAG	GCT	GCT	GCT	1356	
human	Gly	Leu	Arg	Gly	Gly	Ala	Gly	Glu	Pro	Gly	Lys	Asn	Gly	Lys	Gly	Ala	Gly	Glu	Pro	Gly	Pro	Arg	Gly	Glu	Arg	Gly	Glu	Ala	Gly	Glu	452	
mouse	-	Thr	-	-	-	Pro	Ser	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
bovine	-	Gln	-	-	-	Ala	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGT	ATT	TCC	CCA	GGT	GTT	CCA	GGA	GCT	AAA	GGC	GAA	GAT	GGC	AAG	GAT	GGA	TCA	CCT	GGA	GAA	CCT	GGT	GCA	AAT	GGG	CTT	CCA	CCA	CCA	1437	
human	Gly	Ile	Pro	Gly	Val	Pro	Gly	Ala	Lys	Gly	Glu	Asp	Gly	Lys	Asp	Gly	Ser	Pro	Gly	Gly	Glu	Pro	Gly	Ala	Asn	Gly	Leu	Pro	Gly	Pro	479	
mouse	-	Ser	-	-	-	Ile	-	-	Pro	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
bovine	-	Ser	-	-	-	Ile	Ala	-	Pro	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	Thr	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGA	GCT	C	GCA	GAA	AGG	GGT	GCT	AG	C	GGC	TTC	CGA	GGA	CCT	GCT	GGA	CCA	AAT	GGC	ATC	CCA	GGA	GAA	AAG	GGT	CCT	GCT	GCT	GCT	1518	
human	Gly	Ala	Ala	Gly	Glu	Arg	Gly	Ala	Pro	Gly	Phe	Arg	Gly	Pro	Ala	Gly	Pro	Asn	Gly	Ile	Pro	Gly	Glu	Lys	Gly	Pro	Ala	Gly	Pro	Ala	506	
mouse	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGA	GAG	CGT	GGT	GCT	CCA	GGC	CCT	GCA	GGG	CCC	AGA	GGA	GCT	GCT	GGA	GAA	CCT	GGC	AGA	GAT	GGC	GTC	CCT	GGA	GGT	CCA	CCA	CCA	CCA	1599	
human	Gly	Glu	Arg	Gly	Ala	Pro	Gly	Pro	Ala	Gly	Pro	Arg	Gly	Ala	Gly	Glu	Pro	Gly	Arg	Asp	Gly	Val	Pro	Gly	Gly	Pro	Gly	Gly	Gly	Gly	533	
bovine	-	Asp	-	-	-	Gly	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGA	ATG	AGG	GGC	ATG	CCC	GGA	AGT	CCA	GGA	GGA	CCA	GGA	AGT	GAT	GGG	AAA	CCA	GGG	CCT	CCC	GGA	AGT	CAA	GGA	GAA	AGT	AGT	AGT	AGT	1680	
human	Gly	Met	Arg	Gly	Met	Pro	Gly	Ser	Pro	Gly	Gly	Pro	Gly	Ser	Asp	Asn	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
bovine	-	Leu	-	-	-	Ile	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGT	CGA	CCA	GGT	CCT	GGT	CCA	TCT	GGT	CCC	CGA	GGT	CAG	CCT	GGT	GTC	ATG	GGC	TTC	CCC	GGT	CCT	AAA	GGA	AAT	GAT	GAT	GAT	GAT	GAT	1761	
human	Gly	Arg	Pro	Gly	Pro	Pro	Gly	Pro	Ser	Gly	Pro	Arg	Gly	Gln	Pro	Gly	Val	Met	Gly	Phe	Pro	Gly	Pro	Lys	Gly	Asn	Asp	Asp	Asp	Asp	587	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGT	GCT	CCT	GGT	AAG	AAT	GGA	GAA	GGA	GGT	GGC	CCT	GGA	GGA	CCT	GGC	CCT	CAG	GGT	CCT	CCT	GGA	AAG	AAT	GGT	GAA	ACT	ACT	ACT	ACT	1842	
human	Gly	Ala	Pro	Gly	Lys	Asn	Gly	Glu	Arg	Gly	Gly	Pro	Gly	Gly	Pro	Gly	Pro	Gln	Gly	Pro	Pro	Gly	Lys	Asn	Gly	Glu	Thr	Thr	Thr	Thr	614	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGA	CCT	CAA	GGA	CCC	CCA	GGG	CCT	ACT	GGT	CCT	GGT	GGT	GAC	AAA	GGA	GAC	ACA	GGA	CCC	CCT	GGT	CCA	CAA	GGA	TTA	CAA	CAA	CAA	CAA	1923	
human	Gly	Pro	Gln	Gly	Pro	Pro	Gly	Pro	Thr	Pro	Gly	Pro	Gly	Asp	Lys	Gly	Asp	Thr	Gly	Pro	Pro	Gly	Pro	Gln	Gly	Leu	Gln	Gln	Gln	Gln	641	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGC	TTG	CCT	GGT	ACA	GGT	GGT	CCT	CCA	GGA	GAA	AAT	GGA	AAA	CCT	GGG	GAA	CCA	GGT	CCA	AAG	GGT	GAT	GCC	GGT	GCA	CCT	CCT	CCT	CCT	2004	
human	Gly	Leu	Pro	Gly	Thr	Gly	Gly	Pro	Pro	Gly	Glu	Asn	Gly	Lys	Pro	Gly	Glu	Pro	Gly	Pro	Lys	Gly	Asp	Ala	Gly	Ala	Pro	Pro	Pro	Pro	668	
bovine	-	-	-	-	-	Ser	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGA	GCT	CCA	GGA	GGC	AAG	GGT	GAT	GCT	GGT	GCC	CCT	GGT	GAA	CGT	GGA	CCT	CCT	GGA	TTG	GCA	GGG	GCC	CCA	GGA	CTT	AGA	AGA	AGA	AGA	2085	
human	Gly	Ala	Pro	Gly	Lys	Gly	Asp	Ala	Gly	Ala	Pro	Gly	Ala	Pro	Gly	Arg	Gly	Pro	Pro	Gly	Leu	Ala	Gly	Ala	Pro	Gly	Leu	Arg	Arg	Arg	695	
bovine	-	Ile	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGT	GGA	GCT	GGT	CCC	CCT	GGT	CCC	GAA	GGA	GGA	AAG	GGT	GCT	GCT	GGT	CCT	CCT	GGG	CCA	CCT	GGT	GCT	GCT	GGT	ACT	CCT	CCT	CCT	CCT	2166	
human	Gly	Gly	Ala	Gly	Pro	Pro	Gly	Pro	Glu	Gly	Gly	Lys	Gly	Ala	Ala	Gly	Pro	Pro	Gly	Pro	Pro	Gly	Pro	Pro	Gly	Ala	Ala	Gly	Thr	Pro	722	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGT	CTG	CAA	GGA	ATG	CCT	GGA	GAA	AGA	GGA	GGT	CTT	GGA	AGT	CCT	GGT	CCA	AAG	GGT	GAC	AAG	GGT	GAA	CCA	GGC	GGC	CCA	CCA	CCA	CCA	2247	
human	Gly	Leu	Gln	Gly	Met	Pro	Gly	Glu	Arg	Gly	Gly	Leu	Gly	Ser	Pro	Gly	Pro	Lys	Gly	Asp	Lys	Gly	Glu	Pro	Gly	Gly	Gly	Pro	Pro	Pro	749	
bovine	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGT	GCT	GAT	GGT	GTC	CCA	GGG	AAA	GAT	GGC	CCA	AGG	GGT	CCT	ACT	GGT	CCT	ATT	GGT	CCT	CCT	GGC	CCA	GCT	GGC	CAG	CCT	CCT	CCT	CCT	2328	
human	Gly	Ala	Asp	Gly	Val	Pro	Gly	Lys	Asp	Gly	Pro	Arg	Gly	Pro	Thr	Gly	Pro	Ile	Gly	Pro	Pro	Gly	Pro	Ala	Gly	Gln	Pro	Pro	Pro	Pro	776	
bovine	-	Val	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
human	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
mouse	GGA	GAT	AAG	GGT	GAA	GGT	GGT	GCC	CCC	GGA	CTT	CCA	GGT	ATA	GCT	GGA	CCT	CGT	GGT	AGC	CCT	GGT	GAG	AGA	GGT	GAA	ACT	ACT	ACT			

Vol. 260

Table 1. Comparison of the sizes of the various domains of human type I and type III procollagens

Data for type I procollagen are from Kuivaniemi *et al.* (1988), Tromp *et al.* (1988) and Bernard *et al.* (1983a,b).

Domain	Amino acid residues		
	Pro α 1(I)	Pro α 2(I)	Pro α 1(III)
Signal peptide	22	22	24
N-Propeptide	139	57	129
N-Telopeptide	17	11	14
3-Helix	1014	1014	1029
C-Telopeptide	26	15	25
C-Propeptide	246	247	246

Fig. 3. Hydropathy plot for the N-terminal regions of type I and type III procollagen

(a) The human α 1(I) chain, (b) the human α 2(I) chain, (c) the human α 1(III) chain. Amino acids are numbered starting with the first residue of the triple helix numbered as +1. The plot was calculated using the program of James M. Pustell and provided by International Biotechnology (catalogue number 81501-81509).

tripeptide unit. The results here establish that the human α 1(III) chain contains 343 triplets and 1029 amino acids (Table 1). Therefore, it is five triplets and 15 amino acids longer than either the α 1(I) or α 2(I) chain of type I collagen from man and several other species. The triple helices of the bovine and murine α 1(III) chains are probably the same length as in the human, but the data on the amino acid sequences of the bovine chain and the nucleotide sequences of the murine chain are incomplete. Other domains of human type III procollagen were essentially the same as those of human type I procollagen (Table 1).

Comparison of hydropathy plots of the human α 1(III) chain with human α 1(I) and α 2(I) chains indicated that

Fig. 4. Hydropathy plot of type I and type III collagen for the C-terminal region of the protein

Symbols as in Fig. 3; the vertical line indicates the end of the triple-helical domain.

Table 2. Codon usage in the triple-helical domains of type I and type III collagens

Data for type I collagen are from Kuivaniemi *et al.* (1988), Tromp *et al.* (1988) and Bernard *et al.* (1983a,b).

Amino acid	α 1(I)	α 2(I)	α 1(III)	Third base
Glycine	0.50	0.51	0.45	U
	0.28	0.22	0.15	C
	0.18	0.22	0.33	A
	0.03	0.05	0.07	G
Codons examined ...	342	342	362	
Proline (total)	0.60	0.62	0.58	U
	0.38	0.21	0.12	C
	0.02	0.16	0.30	A
	0	0.01	0	G
Codons examined ...	235	199	240	
Proline (Yaa position)	0.84	0.73	0.56	U
	0.13	0.09	0.06	C
	0.03	0.16	0.36	A
	0	0.02	0.02	G
Codons examined ...	117	91	143	
Alanine	0.75	0.76	0.68	U
	0.20	0.15	0.15	C
	0.05	0.08	0.16	A
	0	0	0.01	G
Codons examined ...	118	107	89	

the additional tripeptide units increase the differences between type III collagen and type I collagen at the ends of the molecules (Figs. 3 and 4). The differences in length and in the hydropathy plots are surprising in the light of the continuing uncertainty as to whether or not type III

	TAA	4401
ACC AAA CTC TAT CTG AAA TCC CAA CAA AAA AAA TTT AAC TCC ATA TGT GTT CCT CTT		4458
GTT CTA ATC TTG TCA ACC AGT GCA AGT GAC CGA CAA AAT TCC AGT TAT TTA TTT CCA		4515
	▲▲	*
AAA TGT TTG GAA ACA GTA TAA TTT GAC AAA GAA AAA TGA TAC TTC TCT TTT TTT GCT		4572
	*	
GTT CCA CCA AAT ACA ATT CAA ATG CTT TTT GTT TTA TTT TTT TAC CAA TTC CAA TTT		4629
	*	
CAA AAT GTC TCA ATG GTG CTA TAA TAA ATA AAC TTC AAC ACT CTT TAT GAT AAC AAC		4686
ACT GTG TTA TAT TCT TTG AAT CGT AGC CCA TCT GCA GAG CAA TGA CTG TGC TCA CCA		4743
GTA AAA GAT AAC CTT TCT TTC TGA AAT AGT CAA ATA CGA AAT TAG AAA AGC CCT CCC		4800
TAT TTT AAC TAC CTC AAC TGG TCA GAA ACA CAG ATT GTA TTC TAT GAG TCC CAG AAG		4857
ATG AAA AAA ATT TTA TAC GTT GAT AAA ACT TAT AAA TTT CAT TGA TTA ATC TCC TGG		4914
	*	
AAG ATT GGT TTA AAA AGA AAA GTG TAA TGC AAG AAT TTA AAG AAA TAT TTT TAA AGC		4971
	* *	*
CAC AAT TAT TTT AAT ATT GGA TAT CAA CTG CTT GTA AAG GTG CTC CTC TTT TTT CTT		5028
GTC ATT GCT GGT CAA GAT TAC TAA TAT TTG GGA AGG CTT TAA AGA CGC ATG TTA TGG		5085
TGC TAA TGT ACT TTC ACT TTT AAA CTC TAG ATC AGA ATT GTT GAC TTG CAT TCA GAA		5142
CAT AAA TGC ACA AAA TCT GTA CAT GTC TCC CAT CAG AAA GAT TCA TTG GCA TGC CAC		5199
AGG GAT TCT CCT CCT TCA TCC TGT AAA GGT CAA CAA TAA AAA CCA AAT TAT GGG GCT		5256
▲		
GCT TTT GTC ACA CTA GCA TAG AGA ATG TGT TGA AAT TTA ACT TTG TAA GCT TGT ATG		5313
	▲	
TGG TTG TTG ATC TTT TTT TTC CTT ACA GAC ACC CAT AAT AAA ATA		5358

Fig. 5. Nucleotide sequence of the 3' non-translated region of the human mRNA for type III procollagen

Symbols: underlined bases, differences in nucleotide sequence from previously published sequence for the same region (Mankoo & Dalgleish, 1988); asterisks, insertions of additional bases not found in previously published sequences; vertical arrows, sites at which additional nucleotides were observed in previous cDNA clones; rectangular boxes, sequences of -AATAAA- that are probably signals for polyadenylation.

and type I collagen co-polymerize into the same fibrils (see Fleischmajer *et al.*, 1981).

An additional observation of interest was a variation recognized between the two large cDNA clones (S413 and S31 in Fig. 1) at nucleotide 2638 and therefore at amino acid position 880 (Fig. 2). The variation consisted of a codon of -CTT- in one clone and a codon of -TTT- in the other clone. The variation converted the codon for leucine (Seyer *et al.*, 1980) to a codon for phenylalanine.

An additional observation was that the human se-

quence did not contain a serine residue near the beginning of the *N*-propeptide previously found in amino acid sequencing of bovine type III procollagen (Brandt *et al.*, 1984) and in nucleotide sequencing of the mouse cDNA (Wood *et al.*, 1987). Also, the next three amino acids in the human sequence differed from the murine. The sequence was -Glu-Ala-Val- in the human pro α 1(III) chain. It was -Asn-Val-Asp- in the murine. However, the three amino acids in the human sequence were identical to the bovine.

Conservation of codon usage between human type III collagen and type I collagen

Previous observations established that the codon usage for amino acids in the triple-helical domain of type I collagen from chick (Boedtker *et al.*, 1985) and man (Bernard *et al.*, 1983a,b; Kuivaniemi *et al.*, 1988; Tromp *et al.*, 1988) showed a preference for U in the third position of codons for glycine, proline and alanine. The data here established that human type III procollagen had about the same preference (Table 2). In the comparison of codons for type III with type I collagen, there seemed to be a greater use of A in the third position of codons in type III collagen than in type I. In addition, G was rarely used in the third base of codons for glycine, proline or alanine in either type III or type I collagen.

The nucleotide sequences of the 5' and 3' non-translated region of the cDNA

The nucleotide sequence of the 5' non-translated region demonstrated the presence of three -ATG- start codons for translation noted previously in the mouse (Liau *et al.*, 1985) and chick (Yamada *et al.*, 1983b) cDNAs (positions -15 to -17 and -67 to -69 in Fig. 2). The first two (at -15 to -17 and -67 to -69 in Fig. 2) were followed by in-frame stop codons.

The nucleotide sequence of the 3' non-translated region of the cDNA confirmed previous reports that the human mRNA contained three polyadenylation signals of -AATAAA- (Fig. 5). Several differences from a previously published nucleotide sequence for the 3' non-translated region (Mankoo & Dagleish, 1988) are shown in Fig. 5.

REFERENCES

- Baldwin, C. T., Constantinou, C. D. & Prockop, D. J. (1989) *J. Biol. Chem.* **264**, 3002-3006
- Bernard, M. P., Myers, J. C., Chu, M.-L., Ramirez, F., Eikenberry, E. F. & Prockop, D. J. (1983a) *Biochemistry* **22**, 1139-1145
- Bernard, M. P., Chu, M.-L., Myers, J. C., Ramirez, F., Eikenberry, E. F. & Prockop, D. J. (1983b) *Biochemistry* **22**, 5213-5223
- Boedtker, H., Finer, M. & Aho, S. (1985) *Ann. N.Y. Acad. Sci.* **460**, 85-116
- Brandt, A., Glanville, R. W., Horlein, D., Bruckner, P., Timpl, R., Fietzek, P. P. & Kuhn, K. (1984) *Biochem. J.* **219**, 625-634
- Byers, P. H. & Bonadio, J. F. (1985) in *Genetic and Metabolic Diseases in Pediatrics* (Lloyd, J. & Scriver, C. R., eds.), pp. 50-90, Butterworths, London
- Chu, M.-L., Weil, D., de Wet, W., Bernard, M., Sippola, M. & Ramirez, F. (1985) *J. Biol. Chem.* **260**, 4357-4363
- Fietzek, P. P., Allman, H., Rauterberg, J., Henkel, W., Wachter, E. & Kuhn, K. (1979) *Hoppe-Seyler's Z. Physiol. Chem.* **360**, 809-820
- Fleischmajer, R., Timpl, R., Tuderman, L., Raisher, L., Wiestner, M., Perlsh, J. S. & Graves, P. N. (1981) *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7360-7364
- Glanville, R. W. & Fietzek, P. P. (1976) *FEBS Lett.* **71**, 99-102
- Gubler, U. & Hoffman, B. J. (1983) *Gene* **25**, 263-269
- Kuivaniemi, H., Tromp, G., Chu, M.-L. & Prockop, D. J. (1988) *Biochem. J.* **252**, 633-640
- Liau, G., Mudryj, M. & de Crombrughe, B. (1985) *J. Biol. Chem.* **260**, 3773-3777
- Loidl, H. R., Brinker, J. M., May, M., Pihlajaniemi, T., Morrow, S., Rosenbloom, J. & Myers, J. C. (1984) *Nucleic Acids Res.* **12**, 9383-9394
- Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
- Mankoo, B. S. & Dagleish, R. (1988) *Nucleic Acids Res.* **16**, 2337
- Mayne, R. & Burgeson, R. E. (1987) *Structure and Function of Collagen Types*, Academic Press, New York
- Messing, J. (1983) *Methods Enzymol.* **101**, 20-78
- Miller, E. J. (1988) in *Collagen Vol. 1* (Nimni, M. E., ed.), pp. 140-156, CRC Press, Boca Raton, FL
- Miskulin, M., Dagleish, R., Kluve-Beckerman, B., Rennard, S. J., Tolstoshev, P., Brantly, M. & Crystal, R. G. (1986) *Biochemistry* **25**, 1408-1413
- Prockop, D. J. & Kivirikko, K. I. (1984) *N. Engl. J. Med.* **311**, 376-386
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463-5467
- Seyer, J. M. & Kang, A. H. (1981) *Biochemistry* **20**, 2621-2627
- Seyer, J. M., Mainardi, C. & Kang, A. H. (1980) *Biochemistry* **19**, 1583-1589
- Toman, P. D., Ricca, G. A. & de Crombrughe, B. (1988) *Nucleic Acids Res.* **16**, 7201
- Tromp, G., Kuivaniemi, H., Stacey, A., Shikata, H., Baldwin, C. T., Jaenisch, R. & Prockop, D. J. (1988) *Biochem. J.* **253**, 919-922
- Wood, L., Theriault, N. & Vogeli, G. (1987) *Gene* **61**, 225-230
- Yamada, Y., Kuhn, K. & de Crombrughe, B. (1983a) *Nucleic Acids Res.* **11**, 2733-2744
- Yamada, Y., Mudryj, M. & de Crombrughe, B. (1983b) *J. Biol. Chem.* **258**, 14914-14919
- Yamada, Y., Mudryj, M., Sullivan, M. & de Crombrughe, B. (1983c) *J. Biol. Chem.* **258**, 2758-2761
- Yamada, Y., Liau, G., Mudryj, M., Obici, S. & de Crombrughe, B. (1984) *Nature (London)* **310**, 333-337

Received 31 October 1988/6 January 1989; accepted 9 January 1989