Structure of a full-length cDNA clone for the preprox2(I) chain of human type I procollagen

Comparison with the chicken gene confirms unusual patterns of gene conservation

Helena KUIVANIEMI, Gerard TROMP, Mon-Li CHU and Darwin J. PROCKOP*

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

A cDNA clone from a human placental library was found to consist of an essentially full-length cDNA of 4.6 kb for the prepro $\alpha 2(I)$ chain of type I procollagen. Nucleotide sequencing of the 5'-end of the cDNA provided a sequence of 1617 nucleotide residues and codons for 539 amino acid residues not previously defined. Comparison of the complete structure of the prepro $\alpha 2(I)$ cDNA with previously reported sequences for the chicken pro $\alpha 2(I)$ gene indicated that 83% of 1366 total amino acid residues were conserved. In the α -chain domain 84% of 1014 amino acid residues were conserved. Also, there was conservation of the previously noted preference for U and C in the third position of codons for glycine, proline and alanine. One major difference between the human and the chicken prepro $\alpha 2(I)$ chain was that the human chain contained 21 fewer proline residues, an observation that probably explains why the triple helix of human type I procollagen unfolds at temperatures that are 1-2 °C lower. In parallel experiments, sequencing of intron–exon boundaries for nine exons of genomic subclones confirmed and extended previous observations that the pro $\alpha 2(I)$ gene, like other genes from fibrillar collagens, has an unusual 54-base pattern of exon sizes that is highly conserved through evolution.

INTRODUCTION

Type I collagen provides the fibrous network that maintains the structural integrity of most tissues in vertebrates and in many other multicellular organisms (for reviews see Prockop & Kivirikko, 1984; Cheah, 1985). The protein is a heterotrimer of two $\alpha 1(I)$ and one $\alpha 2(I)$ chains, and it is first synthesized as a procollagen comprised of two $pro\alpha 1(I)$ and one $pro\alpha 2(I)$ chains. Large parts of the primary structure of the $\alpha 1(I)$ and $\alpha 2(I)$ chains of type I collagen are now known. Most of the initial data were generated by Edman degradation of peptide fragments (see Kang et al., 1967; Fietzek et al., 1972; Piez, 1976; Dixit et al., 1978; Hofmann et al., 1978; Galloway, 1982), but, because analysis of the protein is far more difficult, most of the recent data were derived from nucleotide sequencing of DNA clones (Bernard et al., 1983a,b; Boedtker et al., 1985; Dickson et al., 1985). However, it has been difficult to obtain a complete amino acid sequence for an $\alpha(I)$ or pro $\alpha(I)$ chain from a single species. The most complete sequence came from the analysis of a combination of cDNA and genomic clones for the $pro\alpha 2(I)$ chain from chicken, an analysis that lacks only the coding sequences of two exons that contain 36 codons (Boedtker et al., 1985). Also, apparently because of the high G+C content of the relatively long mRNAs, no full-length cDNA is available for a $pro\alpha I(I)$ or $pro\alpha 2(I)$ chain of type I procollagen.

In the present paper we describe the first full-length

cDNA clone coding for a prepro $\alpha 2(I)$ chain. Nucleotide sequencing of the cDNA has permitted detailed comparison of evolutionary differences between the human and chicken procollagen gene. Also, since the full-length cDNA developed here is for the human prepro $\alpha 2(I)$ chain, the clone provides both a probe and structural information that will be of great use in current attempts to define mutations in type I procollagen genes that produce heritable disorders of connective tissue such as osteogenesis imperfecta and Ehlers-Danlos syndrome (see Prockop & Kivirikko, 1984; Byers & Bonadio, 1985; Prockop & Kuivaniemi, 1986).

In the text of the present paper amino acid positions are numbered by the standard convention in which the first glycine residue of the triple-helical domain of an α chain is number 1. The numbers for the $\alpha 2(I)$ chain can be converted into those of the human prepro $\alpha 2(I)$ chain in Fig. 1 by adding 90.

Preliminary reports of this work were presented at the East Coast Connective Tissue Society Meeting, Wood Hole, MA, U.S.A., in March 1987.

MATERIALS AND METHODS

cDNA and genomic clones

A cDNA library from human placenta was obtained from a commercial source (catalogue no. HL 1008; Clontech, Palo Alto, CA, U.S.A.). The library was

^{*} To whom correspondence should be addressed.

These sequence data have been submitted to the EMBL/GenBank Data Libraries under the accession number Y00724.



Fig. 1. Partial restriction map of the cloned cDNA for the preproa2(I) chain (Hp 2010)

Also shown, the relative sizes of two clones previously reported (Bernard *et al.*, 1983*a*) for the human $pro\alpha 2(I)$ chain (Hf-1131 and Hf-32). Symbols: ATG, start site of translation; STH, start of the triple-helical domain; ETH, end of triple-helical domain; TAA, end of translation; A, *AvaI* site; E, *Eco*RI site; M, *MstII* site; N, *NcoI* site; P, *PvuII* site; Ps, *PstI* site; R, *RsaI* site; X, *XhoI* site. The asterisk indicates a variable *PvuII* site that was found in Hp-2010 and in the two clones of genomic allele sequenced here, but not in a short cDNA clone for the pro $\alpha 2$ chain (Hf-15; M.-L. Chu & M. Bernard, unpublished work). As noted in the text, regions not sequenced in both directions were confirmed by analysis of genomic subclones.

prepared with cDNA that was not size-selected, that was inserted into the bacteriophage vector $\lambda gt11$, and that was amplified before distribution. It contained 10⁶ independent clones with an average size of 1.8 kb and a range of sizes of 0.8–3.6 kb. The probe used to screen the library was a 2 kb *Hind*III–*Hind*III fragment that was a subclone of the pro $\alpha 2$ (I) genomic DNA obtained from a human bacteriophage library (clone NJ-3; Myers *et al.*, 1983). The 2 kb *Hind*III–*Hind*III fragment was subcloned into plasmid pBR322 from clone NJ-3 by Mr. Bruce Vogel. One series of genomic subclones contained exons 10-12 of the pro $\alpha 2(I)$ gene and the other contained exons 25-30. The first series was prepared with 10 kb EcoRI-EcoRI fragments of genomic DNA from a proband with atypical Ehlers-Danlos syndrome (Sippola *et al.*, 1984; de Wet *et al.*, 1986). The fragments were cloned into the bacteriophage vector Charon 4A, the clones were screened with a 2 kb *Hind*III-*Hind*III fragment from the pro $\alpha 2(I)$ genomic clone NJ-3, and a 2.3 kb *Bam*HI-EcoRI fragment was subcloned into M13 for nucleotide sequencing. The normal allele was dis-

Fig. 2. Nucleotide and amino acid sequence of the cDNA clone for the proa2(I) chain

The nucleotides are numbered from the start site for transcription (Dickson *et al.*, 1985) and the amino acids from the first amino acid residue of the prepro $\alpha 2(I)$ chain. The 2379 bp nucleotide sequences from the 5'-end of the cDNA clone are indicated in the second line. The overlap with previously published sequences from Hf-32 (Bernard *et al.*, 1983*a*) starts at position 2002. The first 7 bp shown here are from the *Eco*RI linker with which the clone was inserted into $\lambda gt11$ vector. The amino acid sequence encoded for by the clone is indicated in the third line. Top line: nucleotide sequences for the chicken prepro $\alpha 2(I)$ chain where they are known and differ from the human sequence. Lines two and three: nucleotide and amino acid sequences of Hp-2010 defined here. Line four: amino acid sequences encoded for by the chicken prepro $\alpha 2(I)$ cDNA and genomic clones reported previously (Boedtker *et al.*, 1985). Line five: amino acid sequences of the bovine $\alpha 2(I)$ chain that was defined by Edman degradation of peptide fragments (see Galloway, 1982). Symbols; -, identical amino acid; ---, missing nucleotide residues in the human or chicken cDNA; +119 a possible start site for translation that ends in a stop codon after 11 nucleotide residues; +136, start site for translation; vertical lines, beginnings of exons indicated; ψ , cleavage site for signal peptidase; ψ , cleavage site for procollagen N-proteinase; \downarrow , beginning of α chain domain. The coding sequences found in exon 16 from the chicken gene are not known. The amino acid residues are not known (see Galloway, 1982).

	chicken human human	GA	ATT	C CGG	TGC GCG	T GGC	T CAG	CA GTG	G Ata	ATA CCT	G C CCG	AAC CCG	CAC GTG	GT ACC	G CAG	GGG	стс	TGC	A Gac	ACA	AGG	AGT	cIe	+11 CAT	9 GTC	G TAA	CAA GTG	G Cta	GAC	+136 ATG	CTC	141
(chicken																				_			-	-	Ser	Lys			-	-	2
	chicken human human	AGC Ser	TTT Phe	GTG Val	GAT Asp	ACG Thr	CGG Arg	T ACT Thr	TTG Leu	TTG Leu	CTG Leu	C CTT Leu	GCA Ala	GTA Va 1	T ACC Thr	C TTA Leu	A TGC Cys	CTA Leu	GCA Ala	ACA Thr	A TGC CVS		IE2, CA TCT	3 GG TTA	AGT CAA G1n	GAG 61 u	C 6AA 614	T ACT Thr	C GTA Val	666	C G Aga	228
(chicken	-	-	-	-	-	-	Ile	•	•	•	•	•	-	-	Ser	Tyr	-	-	-	Ser	-	His	Val	Ser	-	Ala	Ser	Ala	Gly	-	31
(1 	chicken human human	AAG Lys	E4 GGC G1y	T CCA Pro	AGA GCC Ala	GGA G1y	C GAT Aso	A AGA Arg	G GGA Gly	CCA Pro	AG CGT Arg	GGA G1y	GAA Glu	AGG Arg	E5 GGT G1y	CCA Pro	CCA Pro	T GGC Gly	A CCC Pro	CCA Pro	GGC Gly	AGA Arg	GAT Asp	GGT G1v	GAA Glu	C GAT Asp	C GGT Gly	A CCC Pro	C ACA Thr	T GGC G1v	CCT Pro	318
	chicken	-	-	-	Arg	-	-	Lys	-	•	Gl'n	-	-	-	-	-	-	•	-	-	-	-	-	-	-	-	•	-	Pro		-	61
	chicken human human	A CCT Pro	C GGT G1 v	C CCA Pro	CCT Pro	GGT G1 v	CCT Pro	A CCT Pro	GGC G1v	CCC Pro	CCT Pro	GGT G1v	T CTC Leu	C GGT G1 v	A GGG G1v	E6 T AAC Asn	TTT Phe	GCT Ala	GCT Ala	CAG Gln	TAT	GAT	CCA	TCT GGA	AAA	CG GGA	C GTT Val	AC GGA	T CTT	GGC	C CCT	405
1	chicken bovine	-	-	-	-	-	-	-	-	-	-	-	•	-	-	-	-	•	-	-	-	Glu	Pro	Ser Phe	Asp	Ala Ala	Ala Lys	Asp -	Phe G1y	- -	-	90
	chicken human		T CCA	ATG	E7 T GGC	TTA	ATG	GGA	сст	AGA	GGC	CCA	ССТ	A GGT	GCA	T GCT	GGA	C T GCC	T CCA	E8 T GGC	сст	CT CAA	G GGT	T TTC	CAA.	T	GT CCT	О С 100	GGT	GAG	сст	495
1	h uman chicken povine	61y _	Pro -	Met - -	G1y - -	Leu -	Met -	61y _	Pro -	Arg -	G1y -	Pro -	Pro -	G1y _	A1a _	Ala Ser Ser	G1y -	Ala Pro	Pro -	G1y - -	Pro	Gln Pro	Gly	Phe	Gln	61y -	Pro Val	Ala Pro Pro	Gly	Glu	Pro	120
								E9				_					_									E10						
	chicken human human chicken	GGT Gly	GAA Glu	CCT Pro -	GGT G1y	CAA Gln	A ACT Thr -	GGT G1y	C CCT Pro -	CAG GCA Ala Gln	GGT G1y	C GCT Ala Pro	CGT Arg	GGT Gly	C CCA Pro -	C GCT Ala Pro	T GGC Gly -	CCT Pro -	A CCT Pro	A GGC G1y	AAG Lys	GCT Ala	GGT Gly	GAA Glu	GAT Asp	GGT G1y -	CAC His	CCT Pro	C GGA Gly	AAA Lys	T CCC Pro -	585
1	bovine	-	•	•	-	•	•	-	•	-	-	-	-	-	-	Pro	-	-	•	-	-	-	-	-	•	-	-	-	-	-	-	150
	chicken human	GGA	A Cga	сст	GGT	GAG	G AGA	T GGA	GTT	C STT	T GGA	T CCA	A CAG	E11 GGT	GCT	CGT	GGT	TTC	сст	GGA	ACT	CĊT	T GGA	CG CTT	сст	GGC	T TTC	G AAA	A GGC	ATT	AGG	675
	human chicken	61y -	Arg -	Pro -	61y -	61u -	Arg -	61y -	Val -	Yal Ala	61y -	Pro -	Gln -	61y -	Ala -	Arg -	G1y -	Phe -	Pro -	Gly -	Thr -	Pro -	G1y -	Leu Pro	Pro	61y -	Phe -	Lys -	. G1y -	Ile -	Arg -	
l	DOVINE	- E12	-	•		-	-	-	•	Pro	•	•	-	-	-	-	-	-	-	- E13	-	-	-	-	-	-	-	-	-	-	-	180
	chicken human human	GGA G1y	CAC His	AAT Asn	GGT G1y	CTG Leu	GAT Asp	T GGA Gly	A TTG Leu	C MG Lys	66A 61y	A CAG Gln	T CCC Pro	GGT G1y	GCT Ala	CCT Pro	C GGT Gly	ACC GTG Val	MG Lys	A GGT G1y	GAA Glu	A CCT Pro	A GGT Gly	GCC Ala	CCT Pro	GGT G1y	GAA Glu	AAT Asn	GGA G1y	C ACT Thr	CCA Pro	765
l	chicken bovine	-	-	:	•	-	-	-	-	Thr Thr	-	-	-	-	:	-	-	Thr -	-	-	-	-	-	-	-	-	-	-	-	-	-	210
	hickor			CCT	E14		~	-	~												-		E15						~		~7	055
	chicken human	GGT		ACA	GGA	6CC	CGT	666 61v	CTT	CCT	GGT	GAG	AGA	66A	CGT	GTT	66T	6CC	CCT	66C	CCA	GCT	GGT	6CC	CGT	660	AGT	GAT	66A	AGT	GTG	822
(chicken	- -	-	Pro	-	-	-	- -	-	-	-	-	-	-	-	Ile -	- -	-	-	- -	-	-	-	-	-	-	- -	-	-	-	Ala	240
				-,-	•			E16															•			E17						••••
ļ	chicken human	C GGT	ccc	ACT GTG	A GGT	ССТ	A GCT	GGT	ccc	ATT	GGG	TCT	GCT	GGC	ССТ	CCA	GGC	TTC	CCA	GGT	GCC	CCT	660	ccc	MG	GGT	GAA	ATT	GGA	C GCT	GTT	945
1	human Chicken Dovine	61y - -	Pro - -	Thr -	ыу - -	-	-		-	- 11e	-	-	-	-	-	-	- -	rne -	-	- GIY	-	-	- -	Pro -	Lys -	- -	- - -	Leu	- -	Pro Pro	Ala -	270
			-	••					-	-											TC	107	-	-				~	E18	r		1025
1	nicken human human	GGT Glv	AAC Asn	GCT Ala	GGT Glv	CCT Pro	A GCT Ala	GGT	ccc	GCC	GGT G1v	CCC	CGT	GGT 61v	6AA 61u	A I GTG Val	GGT		CCA	66C	CTC	TCC	GGC	CCC	GTT Val	GGA	CCT	CCT	GGT G1v	AAT Asn	CCT Pro	1035
. 1	chicken bovine	-	-	Glu Pro	-	-	Thr -	-	-	-	-	-	-	-	-	ile -	-	-	-	-	Ser	-	-	-	-	-	-	-	=	-	ĀĨa	300
	chicken	Ŧ	т	T	т		r	6	Ŧ					E19		Ŧ	Ŧ			т		т	т	6		G		т				1125
	human human	GGÅ Gly	GCÅ Ala	AAC Asn	GGC G1v	CTT Leu	ĂCT Thr	GGT G1v	GCC Ala	MG	GGT G1v	GCT	GCT Ala	GGC G1v	CTT Leu	ccc Pro	66C	GTT Val	GCT Ala	666 61 v	GCT Ala	CCC Pro	66C 61y	CTC	CCT Pro	GGA G1y	CCC Pro	CGC Arg	GGT Gly	ATT Ile	CCT Pro	
	chicken bovine	-	Pro	-	-	-	Pro Pro		-	-		•	-	-	-	•	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	330
	chicken	т		~~	c	~				c		•	c				E20				т		т	6	G			т				1215
	human human	GGC G1y	CCT Pro	GTT Val	GGT Gly	GCT Ala	GCC	GGT G1 y	GCT Ala	ACT	GGT Gly	GCC Ala	AGA Arg	GGA G1y	CTT Leu	GTT Val	GGT G1 y	GAG Glu	CCT Pro	GGT G1y	CCA Pro	GCT Ala	GGC Gly	TCC Ser	AAĂ Lys	66A 61 y	GAG Glu	AGC Ser	GGT Gly	AAC Asn	AAG Lys	
	chicken bovine	-	-	Pro-	-	Pro -	-	-	Pro Ala	Ser -	-	-	-	-	-	-	-	-	-	-	-	:	-	Ala -	-	-	-	-	-	-	-	360
	chicken			-	E21					T7	· ~							c		r				AGC		т		ç	C		Ţ	1305
	human human	GGT G1y	GAG Glu	CCC	GGC	TCT Ser	GCT Ala	666 660	CCC Pro	CAA G1n	GGT G1y	CCT Pro	CCT Pro	GGT G1y	CCC Pro	AGT Ser	GGT G1y	GAA Glu	GAA Glu	GGĂ GIY	AAG Lys	AGA Arg	66C 61y	CCT	AAT Asn	GGG G1y	GAA Glu	GCT Ala Pro	GGA Gly	TCT Ser	GCC Ala	
		-						-		0		-								-		-	-	->er	-	-	-	- r i U	-	-		

									1																			1			
chicken human human	T GGC Gly	C CCT Pro	T CCA Pro	C GGA Gly	CCT Pro	G CCT Pro	T GGG G1y	A CTG Leu	AGA Arg	EZZ C GGT G1y	GAG AGT Ser	CCT Pro	A GGT Gly	TCT Ser	CGT Arg	GGT G1y	C CTT Leu	CCT Pro	GGA Gly	GCT Ala	GAT Asp	GGC Gly	AGA Arg	GCT Ala	T GGC Gly	GTC Val	ATG Met	E23 A GGC G1y	CCT Pro	G CCT Pro	1395
chicken bovine	-	- -	-	-	•	Ala -	-	-	-	-	Glu Asn	•	-	•	-	-	-	-	-	-	-	-	-	-	-	-	-		•	Ala Ala	420
chicken human human chicken	GGT G1y	AC AGT Ser Asn	CGT Arg	GGT Gly	T GCA Ala	AGT Ser	A GGC Gly	CCT Pro	T GCT Ala Val	T GGA Gly	CT GTC Val	AAG CGA Arg	T GGA Gly	CCT Pro	AAT Asn	T GGA Gly	GAT Asp	GCT Ala	C GGT Gly	T CGC Arg	CCT Pro	T GGG Gly	A GAG G1u	CCT Pro	GGT Gly	T CTC Leu	ATG Met	T GGA Gly	A CCC Pro	AGA Arg	1485 450
bovine	- E24	-		-	Thr	-	-	-	-	-	-		•	-	-	-	-	Ser	- E 25	•	-	-	-	-	•	-	-	-	-	-	1575
chicken human human chicken	GGT G1y	CTT Leu	CCT Pro -	GGT Gly	TCC Ser Gln	CCT Pro -	GGA G1y	AAT Asn Ser	ATC Ile Pro	GGC Gly	CCC Pro -	GCT Ala	GGA Gly	AAA Lys	GAA Glu	GGT Gly -	CCT Pro	GTC Val	T GGC G1y	T CTC Leu Phe	CCT Pro -	A GGC Gly	GCA ATC 11e Ala	T GAC Asp	T GGC Gly	AGG Arg	GT CCT Pro Val	G GGC Gly	CCA Pro	C ATT Ile	480
bovine	-	Phe	-	-	-	-	-	-	-	-	-	-	-	-	-	-	•	•	-	-	-	- E26	-	-	-	-	-	-	-	-	1665
chicken human human chicken bovine	T GGC Gly	CCA Pro	C GCT Ala	T GGA Gly	AAT GCA Ala Asn Pro	AGA Arg	T GGA Gly	A GAG Glu	CCT Pro	GGC G1y	AAC Asn -	ATT Ile	GGA G1y	TTC Phe	CCT Pro -	GGA G1y	A CCC Pro -	AAA Lys	T GGC Gly -	CCC Pro -	ACT Thr	GGT G1 <i>y</i> -	GAT Asp Glu Glu	CCT Pro -	GGC G1y -	AAA Lys	CCT AAC Asn Pro	GGT G1 <i>y</i> -	A GAT Asp Glu Glu	AM Lys	510
chicken	c		т				c	C A		E27		т	N		6			С				' T			т	GTC	ACT	E28	AAC		1755
human human chicken bovine	GGT Gly	CAT His Asn	GCT Ala Val Val	GGT Gly -	CTT Leu	GCT Ala	GGT Gly	GCT Ala Pro	CGG Arg -	GGT G1y - -	GCT Ala	CCÀ Pro	GGŤ Gly	CCT Pro -	GAT Asp Glu	GGA Gly -	AAC Asn -	AAT Asn	GGT Gly -	GCT Ala	CAG Gln	GGÁ Gly	CCT Pro -	CCT Pro -	GGÁ Gly -	CCA Pro Val Leu	CAG Gln Thr	GGT G1y -	GTT Val Asn	CAA Gln -	540
chicken		C	G			ACA		Ţ	C	C	Å	T		***		E29		Å	T		C		Å			~ • •	C				1845
human human chicken bovine	GGT Gly -	GGA Gly Ala Glu	Lys -	GGT Gly	GAA Glu -	CAG Gln Thr -	GGT Gly -	Pro -	GAT Asp Ala Ala	GGT Gly	Pro	Pro -	GGC Gly	Phe -	CAG Gln -	GGT Gly -	Leu	Pro -	GGC G1y -	Pro -	Ser Ala	GGT Gly - -	CCC Pro -	GCT Ala -	GGT Gly -	GAA Glu -	GTT Val Ala Ala	GGC Gly - -	AAA Lys - -	CCA Pro -	570
chicken	C			E30				A			G					C		G	E31 C	G	T		TT	Ţ		A	C			π	1935
human human chicken bovine	GGA Gly	GAA	AGG	GGT Gly -	Leu	CAT His -	GGT G1y -	GAG Glu	TTT Phe	GGT Gly	CTC Leu Val	CCT Pro	GGT Gly	CCT Pro	GCT Ala	GGT Gly	Pro	AGA Arg	GGG G1y -	GAA Glu	CGC Arg	GGT Gly	CCC Pro Leu	CCA Pro -	GGT Gly -	GAG G1u	AGT	GGT Gly -	GCT Ala -	GCC Ala Val	600
chicken	•	-	6	•	-			-	1	-	-		c	A		c		- 1	י - ד		-	E32	G	-	-	- MT	~'y C	-	c	- ст	2025
human human chicken bovine	66 61: -	T CC y Pri -	T AC D Th Al	T GG r Gl a -	r cci y Pro - -	ITA T Ile -	r GGJ 9 61) -	AGC Sei	cgi Arg	GGT Gly	CCT Pro	TCT Ser	GGA G1y -	CCC Pro -	CCA Pro	GGG Gly -	CCT Pro	GAT Asp -	GGA Gly	AAC Asn -	AAG Lys -	GGT Gly -	GAA Glu -	CCT Pro Ala	GGT Gly -	GTG Val Asn Ala	GTT Val	GGT Gly -	GCT Ala Pro	GTG Val Ala	630
chicken human	66	T G C Ac	C T GC	T GG) T CC/	G /	N 6 1 661		G [Ag1	GGA	A CTC	CCA	GGA	GAG	A AGG	GGT	T	A GCT	T	G T ATA	A CCT	GGA	GGC	AAG	T GGA	G GAA	AAG	E33 GGT	CC GAA	A CCT	2115
human chicken bovine	Gĩ	y Th Al	r Al a Pr	a G1 0 -	y Pro	Sei Ala	G1) G1	y Pro	Ser Glj	Gly	Leu Ile Ile	Pro -	Gly	Glu - -	Arg -	Gly -	Ala Val Val	Ala -	Gly	lle Val Val	Pro -	Gly -	Gly -	Lys - -	Gly -	Glu	Lys - -	Gly	Glu Ala	Pro	660
chicken human	GG	т ст	C AG	A GG	T GA	C C A AT	T 661	GC/ T AAI			AGA	GAT	GGT	C GCT	CGT	E34	CT GCT	C	GGT	GCT	A T GTA	GGT	T GCC	сст	C GGT	сст	GCT	T GGA	GT GCC	G T Aca	2205
human chicken bovine	G1 -	y Le	u Ar	g G1	y G1 As As	u 110 p Thi P	e Gly r -	y Asi Ala	n Pro n Thi	Gly	Arg _	Asp 	Gly	Ala -	Arg	Gly	Ala Leu Leu	His Pro	Gly	Ala -	Val Ile Ile	Gly -	Ala -	Pro -	G1y -	Pro -	Ala -	G1y -	Ala Gly	Thr Ala	690
chicken			T	E3	5 T	G	<u>c</u>	T C	_		C	:					GC	Т	T	Π		E36			Ţ		ССТ			Ţ	2295
human human chicken bovine	GG	iy As	ip Ai	56 66 rg 61	y G1	A GC U A1 G1	T GG a Gly y -	G GC y Ala Pro	T GCI a Ala o -	r GG1 1 G1y -	r CCT / Pro -	GCT Ala -	GGT Gly -	CCT Pro -	GCT Ala -	GGT Gly -	CCT Pro Ala	CGG Arg	GGA Gly -	AGC Ser Ile Ile	CCT Pro - Arg	GGT Gly -	GAA Glu -	CGT Arg -	GGC G1y -	GAG Glu -	GIC Val Pro	GG1 G1y -	Pro -	GCT Ala Val Val	720
chicken human	GG	T ·	T (C N	at NC GG	A TT	T GC	T GG	A T CC	T C G GC1	E37 GGT	A GCT	GCT	GGT	G	T CCG	GGT	GCT	A AA	C GGA	G GAA	C T Aga	N GGA	т о о	G	A GGG	A CCT	AAG	GGT	GAA /	CA	2385
numan chicken bovine	6	iy Pi 	-0 A9 - Se	in Gl ir -	y Ph - -	e A1. - -	a G1; - -	y Pri - -	o Ala Pro	G1) - -	/ A1a _ _	Ala - -	G1y - -	Gln - -	Pro - -	- -	A14 - -	Lys - -	ыу - -	- -	Arg - -	- - -	Pro Thr	- -	- -	- -	Val	- -	-	nsn Thr Gln	750
chicken human human chicken bovine	60 61	A Cl GT G Iy Va - Pi - Pi	CA AO IT GI 1 Va no Ti no Ti	CA IT GO 11 G1 hr -	G IT CC y Pr A1	T T C AC o Th a II G1	T AGG rG1; e - n -	G C C CC y Pro - -	A A C GT o Va Il -	ا لوی ا راہ ا	GCT Ala	T GCT Ala Ser	A GGC G1y	CCA Pro	C A GCI Ala Pro Pro	E38 G															2431 765

tinguished from the abnormal allele by the presence of a 19 bp deletion in the abnormal allele that produced an abnormally spliced mRNA, whereas the normal allele produced mRNA that was normally spliced (H. Kuivaniemi, C. Sabol, G. Tromp, M. Sippola-Thiele & D. J. Prockop, unpublished work). The second series of genomic subclones was prepared with 6 kb HindIII-HindIII fragments of genomic DNA from a proband with a lethal variant of osteogenesis imperfecta (de Wet et al., 1983, 1986). The fragments were cloned into the bacteriophage vector Charon 21A, and the clones were screened with a 3.6 kb HindIII-EcoRI fragment from the $pro\alpha 2(I)$ genomic clone NJ-3. A series of 3.6 kb HindIII-EcoRI fragments from positive clones were subcloned into M13 for nucleotide sequencing. The normal sequences for genomic DNA covering exons 25-30 and their flanking sequences were distinguished from the abnormal ones by the fact that the abnormal allele had a single base mutation and gave rise to an abnormally spliced mRNA (G. Tromp & D. J. Prockop, unpublished work).

Nucleotide sequencing of the cDNA and genomic clones

The cDNA library was first plated on superconfluent cultures with about 100000 clones on each of ten 15 cmdiameter Petri dishes. Twenty-three positive clones were plaque-purified. Fragments of Hp-2010 were subcloned into M13mp18 and M13mp19, and the nucleotide sequencing with the dideoxy method was carried out by using universal primers (Sanger et al., 1977; Messing, 1983). Each fragment was sequenced at least three times. The sequences were first defined by using the Klenow fragment of DNA polymerase I, but most of the sequences were verified by using a modified T7 DNA polymerase (Sequenase, as supplied by U.S. Biochemical Corp.; Tabor & Richardson, 1987) and the dGTP analogue dITP. Use of the T7 DNA polymerase was essential to establish some of the sequences because of the high C+G content of the cDNA. The genomic subclones in M13 were sequenced with the same procedures except that the subclones containing exons 10-12 were sequenced first with primers specific for sequences in the exons and then with primers specific for adjacent regions of the intervening sequences (H. Kuivaniemi, Č. Sabol, G. Tromp, M. Sippola-Thiele & D. J. Prockop, unpublished work).

Other procedures

All the DNA probes were labelled by nick-translation with $[\alpha^{3^2}P]dCTP$. Standard procedures were used for Northern-blot analysis of total RNA from normal human skin fibroblasts and for Southern-blot analysis of genomic fragments.

RESULTS AND DISCUSSION

Restriction map and nucleotide sequence of the cDNA

The cDNA library from human placenta was screened with a 2 kb *Hin*dIII–*Hin*dIII fragment from the 5'-end of the human gene for the $pro\alpha 2(I)$ chain of type I procollagen (Myers *et al.*, 1983). Twenty-three positive clones were plaque-purified and the inserts were analysed by digestion with EcoRI. Four of the clones had inserts ranging from 3.6 to 4.8 kb. By Northern blot analysis, all four cDNAs hybridized (results not shown) to the mRNAs of three different sizes previously demonstrated to be specific for the pro $\alpha 2(I)$ chain of type I procollagen (Myers *et al.*, 1983). The 4.6 kb clone Hp-2010 was selected for detailed analysis. When used as a probe for Southern blot analysis, it hybridized to all the expected fragments from two genomic clones for the pro $\alpha 2(I)$ gene (NJ-1 and NJ-3 in Myers *et al.*, 1983; results not shown).

The 3'-end of the cDNA clone Hp-2010 had the same restriction-endonuclease map as previously reported cDNAs from the 3'-end of the $pro\alpha 2(I)$ gene (Bernard et al., 1983a). These previously analysed cDNAs covered the codons for amino acid residues 533 to 1014 of the α chain domain, the C-terminal telopeptide, all the 243 amino acid residues of the C-propeptide, and the 3'-nontranslated region of the mRNA. In total, they included 2468 bp of the mRNA and codons for 740 amino acid residues. Here we determined 1617 nucleotide residues and the codons for 539 amino acid residues not previously defined (Figs. 1 and 2). In addition, we re-examined the sequence of 425 bp that overlapped the 5'-end of the previously published cDNA sequences (Bernard et al., 1983a) and 327 bp defined by sequencing genomic clones containing exons 1, 4, 5 and 6 (Dickson et al., 1985). As indicated in Fig. 1, over 80% of the base-pairs were sequenced in both directions. The regions not sequenced in both directions were confirmed by nucleotide sequencing of the corresponding exons in genomic subclones (see below).

The data revealed several minor differences from previously reported sequences. In regions analysed in cDNAs (Bernard et al., 1983a) there was a single base difference that converted an alanine codon for amino acid position 653 of the $\alpha 2(I)$ chain into a codon for glycine. In the sequenced exons from the 5'-end of the gene (Dickson et al., 1985), there was a single base difference that converted a proline codon in amino acid position 59 of the prepropeptide into a codon for threonine. These two differences may or may not be variants in the gene structure. Also, the data defined a G and a T residue in the 5'-non-translated region that were previously ambiguous (Dickson et al., 1985). In addition, the data suggested the possible presence of a new polymorphic site (Tsipouras et al., 1983; Grobler-Rabie et al., 1985; Sykes et al., 1986) for cleavage of the gene by PvuII in exon 25 (Fig. 1). The site was present in a previously isolated genomic clone for the $pro\alpha 2(I)$ gene (clone NJ-3; Myers et al., 1983). It was present in subclones of both alleles from the proband with a lethal variant of osteogenesis imperfecta (de Wet et al., 1983). It was found in the DNA clone Hp-2010. However, it was not found in Hf-15, a cDNA clone of about 0.8 kb that was isolated from a human skin fibroblast library (M.-L. Chu & M. Bernard, unpublished work).

Conservation of amino acid sequences

The data as a whole made it possible to make a detailed comparison (Fig. 2 and below) of the nucleotide and amino acid sequences of the human prepro $\alpha 2(I)$ chain and the chicken prepro $\alpha 2(I)$ chain (Boedtker *et al.*, 1985). About 83 % of the 1366 of the amino acid residues in the total human chain were identical with the chicken

H. 1	Kui	vanien	ni	and	others
------	-----	--------	----	-----	--------

Amina said	Amino acid in chicken chain												
in human chain	Ala	Ser	Pro	Val	Th								
Ala		4	18	6	2								
Ser	5		2	1	4								
Pro	8	3		3	2								
Val	9	1	5		4								
Thr	4	3	5	1									

Table 1. Common amino acid substitutions between human and chicken preproα2(I) chains

Changes indicated account for 90 out of 227 amino acid substitutions

chain. In the α -chain domain 84% of the 1014 amino acids residues were identical.

Most of the amino acid differences were conservative substitutions. Ninety (40%) of 227 differences in the prepro α 2(I) chains were accounted for by five residues: alanine, serine, proline, valine and threonine (Table 1). The most striking difference in amino acids between the human and the chicken chain was that the human chain contained 21 fewer proline residues in the triple-helical domain. Of these, 14 were in the Yaa-position and therefore are likely to be converted into 4-hydroxyproline during procollagen biosynthesis (see Prockop & Kivirikko, 1984). The terminal stability of the collagen triple helix depends in large part on its content of proline and hydroxproline residues. Therefore the difference in imino acids probably explains why the triple helix of the chicken type I procollagen unfolds with a T_m ('melting' temperature) of about 42 °C whereas human type I procollagen has a T_m that is 1–2 °C lower (Hayashi *et al.*, 1979; Peltonen *et al.*, 1980).

Conservation of codon usage between the human and chicken $pro\alpha 2(I)$ cDNAs

Previous reports indicated an unusual preference for U or C in the third base of codons for glycine, proline and alanine in cDNAs for human and chicken $\alpha 1(I)$ and $\alpha 2(I)$ chains (Bernard *et al.*, 1983*a,b*; Boedtker *et al.*, 1985). As indicated in Table 2, the same preference was largely maintained when the data for the whole human $\alpha 2(I)$ chain were analysed. Of note was that there was a greater preference in the human $\alpha 2(I)$ sequence for U in the third position of codons for proline that were in the Yaaposition of the repeating -Gly-Xaa-Yaa- sequence of the collagen α -helix than in the total proline codons. The preference for U in the third position of the Yaa-position proline codons is probably explained by the avoidance of



Fig. 3. Nucleotide sequences of the exon/intron boundaries of nine exons and the lariat-loop sites for the 5' intervening sequences

Symbols: +, nucleotides that are identical with the consensus sequence; -, nucleotides that lack identity with the consensus sequence; numbers in parentheses, number of bases between those shown. The normal consensus sequence is PyNPyTPuAPy (Ruskin *et al.*, 1984).

Table 2. Codon usage in $\alpha 2(I)$ domain of chicken and human type I procollagen

Data for the chicken $\alpha 2(I)$ chain are from Boedtker *et al.* (1985). No nucleotide sequence data are available for exons 16 and 24 from the chicken gene. These exons contain 36 amino acid residues in the human pro $\alpha 2(I)$ gene and include ten proline, 12 glycine and three alanine residues (see Fig. 2).

Amino acid	Chick	Human	Third base
Gly	0.64 0.18 0.16 0.02	0.51 0.22 0.22 0.05	U C A G
Pro (total)	0.70 0.09 0.20 0.01 211	0.62 0.21 0.16 0.01	U C A G
Pro (Yaa-position) Codons examined	0.81 0.02 0.17 0 100	0.73 0.09 0.16 0.02 91	U C A G
Ala Codons examined	0.82 0.07 0.10 0 97	0.76 0.15 0.08 0 107	U C A G

C-G dinucleotide sequences (Bird, 1986; Brown & Bird, 1986) when the Yaa-position proline is followed by glycine (GGN).

Conservative 54 bp pattern in exons of $pro\alpha 2(I)$ gene

In parallel experiments, nucleotide sequencing was carried out on two series of genomic subclones for the human $pro\alpha 2(I)$ gene. Data were generated for nine exons not previously sequenced (Fig. 3). Seven of the exons were 54 bp, one was 45 bp and one was 99 bp. These exons were previously reported to have the same sizes in the chicken $pro\alpha 2(I)$ gene (Boedtker *et al.*, 1985). Therefore the results extend previous observations indicating that the $pro\alpha 2(I)$ gene, like other genes for fibrillar collagens (Ohkubo et al., 1980; Vogeli et al., 1980; Yamada et al., 1980; Dickson et al., 1985; Boedtker et al. 1985; Cheah, 1985), has an unusual 54 bp pattern of exon sizes and that the pattern is highly conserved through evolution. The data presented in Fig. 3 also provide the nucleotide sequences of the exon/intron boundaries and the sites for lariat-loop formation in nine of the intervening sequences. The boundary sequences and the sites for the lariat loop are, in general, homologous with similar sequences in other genes (Mount, 1982; Ruskin et al., 1984; Reed & Maniatis, 1985; Padgett et al., 1986).

The work presented here was supported in part by National Institutes of Health Research Grant AM-38188 and by a grant from the March of Dimes-Birth Defects Foundation.

REFERENCES

- 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
- Bird, A. P. (1986) Nature (London) 321, 209-213
- Boedtker, H., Finer, M. & Aho, S. (1985) Ann. N.Y. Acad. Sci. 460, 85–116
- Brown, W. R. A. & Bird, A. P. (1986) Nature (London) 322, 477-481
- Byers, P. H. & Bonadio, J. F. (1985) in Genetic and Metabolic Diseases in Pediatrics (Lloyd, J. & Scriver, C. R., eds.), pp. 56–90, Butterworths, London
- Cheah, K. S. E. (1985) Biochem. J. 229, 287-303
- de Wet, W. J., Pihlajaniemi, T., Myers, J., Kelly, T. E. & Prockop, D. J. (1983) J. Biol. Chem. 258, 7721-7728
- de Wet, W., Sippola, M., Tromp, G., Prockop, D., Chu, M.-L. & Ramirez, F. (1986) J. Biol. Chem. 261, 3857-3862
- Dickson, L. A., de Wet, W., Di Liberto, M., Weil, D. & Ramirez, F. (1985) Nucleic Acids Res. 13, 3427–3438
- Dixit, S. N., Seyer, J. M., Kang, A. H. & Gross, J. (1978) Biochemistry 17, 5719-5722
- Fietzek, P. P., Wendt, P., Kell, I. & Kuhn, K. (1972) FEBS Lett. 26, 74–76
- Galloway. D. (1982) in Collagen in Health and Disease (Weiss, J. B. & Jayson, M. I. V., eds.), pp. 528-557, Churchill-Livingstone, Edinburgh
- Grobler-Rabie, A. F., Wallis, G., Brebner, D. K., Beighton, P., Bester, A. J. & Mathew, C. G. (1985) EMBO J. 4, 1745–1748
- Hayashi, T., Curran-Patel, S. & Prockop, D. J. (1979) Biochemistry 18, 4182–4187
- Hofmann, H., Fietzek, P. P. & Kuhn, K. (1978) J. Mol. Biol. 125, 137-165
- Kang, A. H., Bornstein, P. & Piez, K. A. (1967) Biochemistry 6, 788-795
- Messing, J. (1983) Methods Enzymol. 101, 20-78
- Mount, S. M. (1982) Nucleic Acids Res. 10, 459-472
- Myers, J. C., Dickson, L. A., de Wet, W., Bernard, M. P., Chu,
 M.-L., Di Liberto, M., Pepe, G., Sangiorgi, F. O. & Ramirez,
 F. (1983) J. Biol. Chem. 258, 10128–10135
- Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V. E., Sullivan, M., Pastan, I. & de Crombrugghe, B. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 7059-7063
- Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S.
- & Sharp, P. A. (1986) Annu. Rev. Biochem. 55, 1119–1150 Peltonen, L., Palotie, A., Hayashi, T. & Prockop, D. J. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 162–166
- Piez, K. A. (1976) in Biochemistry of Collagen (Ramachandran, G. N. & Reddi, A. H., eds.), pp. 1–44, Plenum Press, New York
- Prockop, D. J. & Kivirikko, K. I. (1984) N. Engl. J. Med. 311, 376–386
- Prockop, D. J. & Kuivaniemi, H. (1986) Rheumatology 10, 246-271
- Reed, R. & Maniatis, T. (1985) Cell 41, 95-105
- Ruskin, B., Krainer, A. R., Maniatis, T. & Green, M. R. (1984) Cell **38**, 317–331
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5463–5467
- Sippola, M., Kaffe, S. & Prockop, D. J. (1984) J. Biol. Chem. 259, 14094–14100

- Sykes, B., Wordsworth, P., Ogilvie, D., Anderson, J. & Jones, N. (1986) Lancet ii, 69-72
- Tabor, S. & Richardson, C. C. (1987) Proc. Natl. Acad. Sci. U.S.A. 84, 4767-4771
- Tsipouras, P., Myers, J. C., Ramirez, F. & Prockop, D. J. (1983) J. Clin. Invest. **72**, 1262–1267
- Vogeli, G., Ohkubo, H., Avvedimento, V. E., Sullivan, M., Yamada, Y., Mudryj, M., Pastan, I. & de Crombrugghe, B. (1980) Cold Spring Harbor Symp. Quant. Biol. **45**, 777–783
- Yamada, Y., Avvedimento, V. E., Mudryj, M., Ohkubo, H., Vogeli, G., Irani, M., Pastan, I. & de Crombrugghe, B. (1980) Cell **22**, 887–892

Received 21 September 1987/23 November 1987; accepted 5 February 1988