Structure of a full-length cDNA clone for the prepro α 1(I) chain of human type I procollagen

Gerard TROMP,* Helena KUIVANIEMI,* Alex STACEY,† Hideo SHIKATA,* Clinton T. BALDWIN,* Rudolf JAENISCH† 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., and †Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, U.S.A.

A full-length cDNA clone for the human prepro α 1(I) chain of type I procollagen was characterized. Nucleotide sequencing of the first 1500 nucleotide residues of the 5'-end of the cDNA clone provided 729 nucleotide residues and the codons for 243 amino acid residues not previously defined from any species. The data made it possible, for the first time, to compare completely codon usage for the human α 1(I) and α 2(I) chains.

INTRODUCTION

Type I collagen is the major component of bone, tendon, skin and dentine (for reviews see Prockop & Kivirikko, 1984; Burgeson & Morris, 1986). It is a heterotrimer with the chain composition [α 1(I)] $_2$ α 2(I), and it is first synthesized as a procollagen comprised of two pro α 1(I) and one pro α 2(I) chains. Although data on the complex structure of type I collagen are now available, a complete nucleotide sequence for the prepro α 1(I) cDNA has not hitherto been determined.

Here we report the partial sequence of a full-length human cDNA clone coding for a prepro α 1(I) chain isolated previously (Stacey *et al.*, 1987). The information

provided by the sequencing of this clone will be of great help in studying mutations in type I procollagen genes in diseases such as osteogenesis imperfecta, Ehlers–Danlos syndrome and Marfan syndrome (see Prockop & Kivirikko, 1984; Byers & Bonadio, 1985; Prockop & Kuivaniemi, 1986).

MATERIALS AND METHODS

Nucleotide sequencing of cDNA clone for prepro α 1(I) collagen

The isolation and partial characterization of the full-length human prepro α 1(I) cDNA clone (pHUCI) has

Fig. 1. Sequencing strategy used and partial restriction map of the cDNA clone for prepro α 1(I) chain (pHUCI)

(a) Sequence strategy used for the about 1.5 kb *EcoRI*/*XhoI* fragment of pHUCI. Arrows starting with X indicate clones that were generated by the Sequenest transposon-deletion system. Other arrows indicate clones starting at corresponding restriction sites. (b) Partial restriction map for entire pHUCI. □, Untranslated region. The regions coding for the protein domains are indicated by shading and hatching: ▨, propeptides; ▩, telopeptides; ▧, triple-helical domain; ▤, signal peptide. Also shown are the relative sizes of two clones previously reported (Chu *et al.*, 1982; Bernard *et al.*, 1983b) for the human pro α 1(I) chain (Hf404 and Hf677). Symbols: Av, *AvaI* site; Ba, *BamHI* site; E, *EcoRI* site; P, *PvuII* site; R, *RsaI* site; X, *XhoI* site.

† To whom correspondence should be addressed.

These sequence data have been submitted to the EMBL/GenBank Libraries.

1988

been described previously (Stacey *et al.*, 1987). For sequencing, the clone was inserted into the Sequenest transposon-deletion vectors called pAA-PZ618 and pAA-PZ619 (Gold BioTechnology, St. Louis, MO, U.S.A.; Peng & Wu, 1986). Deletions were generated according to manufacturer's recommendations. Single-stranded templates for sequencing were obtained by single-strand rescue by use of the helper bacteriophage M13K07 and sequenced by using a primer specific for the vectors pAA-pZ618 and pAA-pZ619. Subsequently, fragments of pHUCI were subcloned into bacteriophages M13mp18 and M13mp19, and sequenced by using universal primers (Sanger *et al.*, 1977; Messing, 1983).

RESULTS AND DISCUSSION

Restriction map and nucleotide sequence of the cDNA

The clone pHUCI (human collagen) is a cDNA clone corresponding to the complete 4.8 kb human pro α 1(I) mRNA (Stacey *et al.*, 1987) (Fig. 1). To establish that the clone was full length and coded for a functional α 1(I) collagen mRNA, it was cloned into a retroviral vector capable of expressing inserted sequences in host mouse cells. Infection of Mov13 mouse cells (Harbers *et al.*, 1984; Schnieke *et al.*, 1987) with the vector resulted in the production of stable type I collagen consisting of two human α 1 chains and one mouse α 2 chain (Stacey *et al.*, 1987).

The previously analysed cDNAs covered the codons for amino acid residues 247–1014 of the α -chain domain, the 26 amino acid residues of the C-terminal telopeptide, the 246 amino acid residues of the C-propeptide and

Table 1. Codon usage in the triple-helical domain of the pro α 1(I) and pro α 2(I) human type I procollagen

Data for the triple-helical domain of the pro α 1(I) chain are from Kuivaniemi *et al.* (1988) and Bernard *et al.* (1983a).

Amino acid	α 1(I)	α 2(I)	Third base
Gly	0.50	0.51	U
	0.28	0.22	C
	0.18	0.22	A
	0.03	0.05	G
Codons examined	342	342	
Pro (total)	0.60	0.62	U
	0.38	0.21	C
	0.02	0.16	A
	0	0.01	G
Codons examined	235	199	
Pro (Yaa position)	0.84	0.73	U
	0.13	0.09	C
	0.03	0.16	A
	0	0.02	G
Codons examined	117	91	
Ala	0.75	0.76	U
	0.20	0.15	C
	0.05	0.08	A
	0	0	G
Codons examined	118	107	

Fig. 2. Nucleotide and amino acid sequence of the cDNA clone for the prepro α 1(I) chain (pHUCI)

The nucleotides are numbered from the start site for transcription (Chu *et al.*, 1985) and the amino acids from the first amino acid of the prepro α 1(I) chain. The 1500 bp nucleotide sequences from the 5'-end of the human cDNA clone are indicated in the second line. The overlap with previously published sequences from Hf404 (Bernard *et al.*, 1983b) starts at nucleotide position 1392 (amino acid residue 425). The amino acid sequence encoded for by the clone is indicated below the nucleotide sequence. Line 1: nucleotide sequences for the chicken prepro α 1(I) chain where they are known (bases 27–571; Finer *et al.*, 1987) and differ from the human sequence. Lines 2 and 3: nucleotide and amino acid sequences of pHUCI defined in the present work. Line 4: amino acid sequences of the chicken prepro α 1(I) chain where they are known, these being residues 1–151 derived from genomic clone (Finer *et al.*, 1987) and residues 158–301 defined by Edman degradation of peptide fragments (see Galloway, 1982). Line 5: amino acid sequence of the bovine pro α 1(I) chain that was defined by Edman degradation of peptide fragments (see Galloway, 1982). Symbols: -, identical amino acid; ---, missing nucleotides in the human or chicken cDNA; *, missing amino acid; empty space in amino acid sequence, not known; +61 and +103, possible start sites for translation that both end in a stop codon after 12 nucleotide residues; +120, start site for translation; vertical lines, beginnings of exons indicated, where known (Chu *et al.*, 1984); ▼, cleavage site for signal peptidase; ↓, cleavage site for procollagen N-proteinase; ↓, beginning of α -chain domain.

224 bp of the 3'-non-translated region of the mRNA. In total, they included 3344 bp of the mRNA and codons for 1040 amino acid residues (Chu *et al.*, 1982; Bernard *et al.*, 1983b). In the present work we determined 729 nucleotide residues and the codons for 243 amino acids residues not previously defined (residues 182–424 in Fig. 2) from any species. In addition, we re-examined the sequence of 626 bp defined by sequencing genomic clones containing exons 1 to 6 (Chu *et al.*, 1984, 1985). As indicated in Fig. 1, the entire 1500 bp was sequenced in both directions.

Comparison of codon usage between the human α 1(I) and α 2(I) cDNAs

The data developed here made it possible, for the first time, to compare completely codon usage for the human α 1(I) chain with that for the human α 2(I) chain. The results (Table 1) demonstrated a marked preference for U as a third base in codons for glycine, proline and alanine, in that U was used for 0.50–0.84 of these codons in both chains. Even more striking was the rare use of G for the third base of the same codons. G was found in only nine of 342 glycine codons in the α 1(I) chain. It was not used for any of the total 353 proline and alanine codons. The results therefore indicate a strong selective pressure against the use of G in these coding sequences. There was also a preference for U in the third position of codons for proline that were in the Yaa position of the repeating -Gly-Xaa-Yaa- sequence of the collagen α -chains. The preference had been noted in the sequence of the α 2(I) chain (Kuivaniemi *et al.*, 1988), but was even more pronounced in the sequence of the α 1(I) chain.

The work presented here was supported in part by N.I.H. Research Grant AR-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. (1983*a*) *Biochemistry* **22**, 1139-1145
- Bernard, M. P., Chu, M.-L., Myers, J. C., Ramirez, F., Eikenberry, E. F. & Prockop, D. J. (1983*b*) *Biochemistry* **22**, 5213-5223
- Burgeson, R. E. & Morris, N. P. (1986) in *Connective Tissue Disease: Molecular Pathology of the Extracellular Matrix* (Uitto, J. & Perejda, A. J., eds.), pp. 3-28, Marcel Dekker, New York
- 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
- Chu, M.-L., Myers, J. C., Bernard, M. P., Ding, J.-F. & Ramirez, F. (1982) *Nucleic Acids Res.* **10**, 5925-5934
- Chu, M.-L., de Wet, W., Bernard, M., Ding, J.-F., Morabito, M., Myers, J., Williams, C. & Ramirez, F. (1984) *Nature (London)* **310**, 337-340
- Chu, M.-L., de Wet, W., Bernard, M. & Ramirez, F. (1985) *J. Biol. Chem.* **260**, 2315-2320
- Finer, M. H., Aho, S., Gerstenfeld, L. C., Boedtker, H. & Doty, P. (1987) *J. Biol. Chem.* **262**, 13323-13332
- Galloway, D. (1982) in *Collagen in Health and Disease* (Weiss, J. B. & Jayson, M. I. V., eds.), pp. 528-557, Churchill-Livingstone, Edinburgh
- Harbers, K., Kuehn, M., Delius, H. & Jaenisch, R. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1504-1508
- Kuivaniemi, H., Tromp, G., Chu, M.-L. & Prockop, D. J. (1988) *Biochem. J.* **252**, 633-640
- Messing, J. (1983) *Methods Enzymol.* **101**, 20-78
- Peng, Z.-G. & Wu, R. (1986) *Gene* **45**, 247-252
- Prockop, D. J. & Kivirikko, K. I. (1984) *N. Engl. J. Med.* **311**, 376-386
- Prockop, D. J. & Kuivaniemi, H. (1986) *Rheumatology* **10**, 246-271
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463-5467
- Schnieke, A., Dziadek, M., Bateman, J., Mascara, T., Harbers, K., Gelinas, R. & Jaenisch, R. (1987) *Proc. Natl. Acad. Sci. U.S.A.* **84**, 764-768
- Stacey, A., Mulligan, R. & Jaenisch, R. (1987) *J. Virol.* **61**, 2549-2554

Received 5 April 1988/23 May 1988; accepted 1 June 1988