Analysis of the promoter region and the N-propeptide domain of the human pro α 2(I) collagen gene

Leon A. Dickson¹, Wouter de Wet², Maurizio Di Liberto^{1,3}, Dominique Weil³ and Francesco Ramirez³

¹Dept. Biochem., Univ. Med. & Dentistry New Jersey, Sch. Osteopath. Med., Piscataway, NJ 08854, USA, ²Dept. Biochem., Univ. Potchefstroom, Potchefstroom, South Africa, and ³Dept. Obstet. Gynecol., Univ. Med. & Dentistry New Jersey, Rutgers Med. Sch., Piscataway, NJ 08854, USA

Received 14 March 1985; Accepted 23 April 1985

ABSTRACT

We have located the exon coding for the start site of transcription of the human pro α 2(I) collagen gene. Comparison with the homologous region of other fibrillar collagen genes has confirmed the existence of a consensus sequence (CATGTCTA-n-TAGACATG) capable of forming a hairpin secondary structure possibly involved in the regulation of collagen biosynthesis. Sequence comparison of the chromosomal regions at the 5' end of the pro α 1(I) and pro α 2(I) collagen genes failed to identify unique DNA elements potentially mediating common regulatory signals. Sequencing of four exons coding for the N-terminal propeptide has determined most of its structure and it has implied the existence of smaller coding units similar to the 11 and 18 bp exons originally described in the avian gene.

INTRODUCTION

Type I collagen is the most abundant among the more than ten different types of collagen identified to date. This trimeric protein is found mainly in skin, bone, tendon and ligaments and it is composed of two pro α 1(I) chains and one pro α 2(I) chain (1). Alterations in either the production or the structure of the two subunits of the type I collagen have been described in some forms of osteogenesis imperfecta, a heterogeneous group of inherited disorders characterized by bone fragility (2). Inhibition of type I collagen synthesis, mediated at the level of transcription, has been reported in chemically and virally transformed cells (3,4). Similarly, the presence of a mutation affecting pro α 2(I) gene expression has been invoked to explain the lethal phenotype of a compound heterozygote affected by osteogenesis imperfecta (5).

To study some of the control elements regulating type I collagen gene expression in physiological and pathological conditions, a number of investigators have analyzed in detail the structure of the 5' portion of the pro α 1(I) and pro α 2(I) collagen gene (6-13). In this paper, we present the identification and characterization of the 5' end of the human pro α 2(I) collagen gene, which includes the promoter region and the exons coding for

the N-propeptide domain. These data have been compared with those obtained from the homologous and cognate collagen genes from chicken, mouse and man in order to identify evolutionarily conserved regions, suggestive of collagen-specific functional domains.

MATERIALS AND METHODS

Genomic Clones

The chromosomal clones used in these studies, which will be described in more detail elsewhere (de Wet et al., in prep.), were isolated from Charon 4A libraries carrying 15-20 kb insert of partially Eco RI digested human DNA. The libraries were kindly provided to us by Drs. A. Bank (Columbia University) and M. Goldsmith (National Institute of Health). To locate the 5' end of the human pro α 2(I) gene, we used as a hybridization probe a mouse genomic fragment 850 bp long which contains the first exon and part of the first intron of the rodent gene (a generous gift of Dr. B. de Crombrughe) (8).

RNA purification

Total poly(A⁺) RNA was isolated from normal human fibroblasts as previously described (13).

Sequencing, S1 nuclease mapping and Nucleic Acid hybridization

DNA sequencing was carried out according to the chemical modification procedure described by Maxam and Gilbert (14). Sequencing of both strands was performed for the regions detailed in this paper. S1 protection experiments of end labeled DNA fragments were carried out according to the published procedures (15). Paper bound nucleic acids hybridizations were performed as previously described (13).

RESULTS

Identification of the 5' end of the pro α 2(I) collagen gene

We have previously reported the isolation of three overlapping clones containing more than 26 kb of the pro α 2(I) collagen gene (16). More recently, by walking on the chromosome, we isolated a series of overlapping clones covering an additional 22 kb of chromosomal DNA. In these studies, (de Wet et al., in prep.) we defined the overall structure of the pro α 2(I) gene which was found to be 35 kb long. More precisely, R-looping analysis in conjunction with Northern blotting hybridizations of the appropriate genomic subclones located, in first approximation, the 5' end of the gene within a 4.0 kb EcoRI chromosomal fragment. This fragment was isolated and subcloned into the EcoRI site of pBR322. The resulting recombinant plasmid, pHA2-1, was

Fig. 1. Restriction map of the 5' end segment of the human pro α 2(I) collagen gene. The sites shown refer to the following enzymes: BI: Bgl I; BII: Bgl II; E: Eco RI; H: Hind III; Hf: Hinf I; P: Pst I; Pv: Pvu II; S: Sph I; X: Xba I; Xh: Xho I. The numbers immediately beneath refer to the size of the EcoRI and Hind III fragments expressed in kilobases. The thickening of the line signifies the regions which have been sequenced. Expanded below the main map is shown a more detailed representation of the region around the beginning of transcription found in pHA2-1. Indicated is the Hinf I fragment used for the S1 protection experiment (arrow) and the relative position of exon 52 (black box). Similarly, but not expanded, are the relative positions of the other three N-terminal domain exons numbered at 49, 48 and 47.

subjected to restriction endonuclease mapping and sequencing analysis. Shown in figure 1 is a partial restriction map of pHA2-1 and of the chromosomal region located immediately 3' to it. In the same figure are outlined two contiguous Hind III genomic fragments (1.6 kb and 4.5 kb) whose analysis is discussed later in the text.

In order to confirm the visual evidence obtained by electron microscopy and to refine the location of the end of the gene within pHA2-1, we used a mouse clone specific for the homologous region of the rodent gene. The mouse probe (pAZ1002) is an 850 bp long Xho I fragment subcloned in pUC8 and it contains part of the pro α 2(I) promoter region, the first exon and part of the first intron (8). Southern blotting hybridization of pAZ1002 to human pHA2-1 DNA digested with various restriction enzymes identified a 240 bp Sph I:EcoRI fragment as the region possibly containing the last exon, number 52 (17), of the human gene (Fig. 1). Sequencing from the EcoRI site identified a 70 bp long open-reading frame coding for the first 23 1/3 amino acid residues of the N-terminal prepropeptide of the pro α 2(I) collagen chain (Fig. 2) (17). To define with more precision the start site of transcription, S1 mapping was employed. A 368 bp Hinf I fragment of pHA2-1 (Fig. 1) was 5' labeled, strand

Fig. 2. DNA and amino acid sequences of four exons coding for the pro $\alpha 2(I)$ N-prepropeptide and their comparison with the homologous sequences from chicken and mouse. The latter sequences have been determined only for exon 52, whereas sequences of the former include those derived from exons 51 and 50 not yet identified in the human gene. Shown are only the differences among the three species. The dots indicate deletions, the triangles demarcate the exons, and the dashes are undetermined sequences. The two arrows indicate the cleavage sites of the signal and N-terminal peptidases (19,21).

3430

Fig. 3. Determination of the site of the initiation of transcription of the human $\text{pro}\alpha 2(\text{I})$ collagen gene. On each side of the five Maxam-Gilbert sequence lanes (G, G+A, A+C, C+T, and C) of the anti-sense strand of the 368 bp Hinf I fragment (Fig. 1) appears the full length fragment (lane -S1, + poly(A⁺) RNA), the protected product after S1 digestion (lane +S1, + poly(A⁺) RNA), and the digested fragment after S1 treatment in the absence of poly(A⁺) RNA (lane +S1, - poly(A⁺) RNA). The sequences of the anti-sense strand (-) and the corresponding complementary sequences of the sense strand (+) are shown for the regions of the initiation of transcription, AGC, the TATA box, and the CAT box.

(6-8, 11,12). The first AUG (+55) is followed by a short open-reading frame potentially coding for five amino acids and ending with the termination signal UGA. Likewise, the second AUG (+119) is followed by the codons for Ser-Lys-Cys and by a second termination signal (UAG). Finally, the third AUG is at position +136 and it is followed by a 70 bp long open-reading frame ending with the intron donor sequence, GTAAGT.

Fig. 4. Nucleotide sequence of the region around the last exon of the *proa2(I)* gene. The triangles locate the position of some of the restriction sites shown in Fig. 1. They are *Ava* I (A), *Hinf* I (Hf), *Pvu* II (P) and *Sph* I (S). Underlined are the CAT and TATA boxes and the start site of transcription (+1). The consensus sequences around the real site of translation are boxed. The vertical line indicates the exon/intron junction. The differences between the human and the rodent sequences are shown.

After the first observation of Vogeli et al. (6), several investigators have reported that around the third AUG of all but one of the fibrillar collagen genes is a uniquely conserved sequence capable of forming a hairpin secondary structure (7,8,12,19). Now we confirmed this finding in the human pro α 2(I) gene (Fig. 5).

Tate et al. defined for the first time the primary structure of the signal peptide and N-terminal peptide (p α 2(I)) of the chick pro α 2(I) chain as well as the arrangement of the exons coding for this domain of the collagen molecule (17). These studies revealed that six exons, including the smallest ever described in eucaryotes, code for the p α 2(I), the N-terminal telopeptide and the signal peptide (19). Interestingly, a similar exon arrangement has been found in the pro α 1(I) gene, although the coding capacity of this segment is twice as large as that of the pro α 2(I) gene (13). These data strongly suggested that although the two small pro α 2(I) exons may represent vestiges of the larger globular domain absent in the p α 2(I) peptide, their persistence might be the result of some, still undetermined, evolutionary pressure. We have sequenced only four of the six exons coding for

Fig. 5. Proposed secondary structure around the initiation site of translation of the type I collagen genes. Enclosed in the dashed box are the consensus sequences.

the human p α 2(I) because, thus far, we have been unable to locate the 11 and 18 bp exons (Fig. 2). Pending the identification of these two small exons, we can safely assume their existence in the human gene because of the remarkable conservation of structural complexity observed in the collagen genes and because exon 52 terminates with a split codon.

DISCUSSION

Analysis of the promoter region and the translation initiation site

Matrix analysis (20) of the region surrounding the 5' end region of the pro α 2(I) gene showed an extremely high level of sequence conservation between man and mouse (Fig. 6A). More precisely, in the 340 bp extending 5' from the start site of transcription the two mammalian genes share a 91% of sequence homology. An almost identical value of homology was observed in a similar interspecies comparison of the pro α 1(I) gene (12). Conserved are also the relative positions of the TATA and CAT boxes as well as the location of the G/C rich segments which in the mouse and chick pro α 2(I) gene have been shown to be a dominant S1 sensitive site (9,10). Using supercoiled plasmids, Finer

Fig. 6. Matrix analysis comparing the pro α 2(I) 5' flanking sequences of man, horizontal axis, and mouse, vertical axis, (Panel A) and the same regions of the pro α 1(I), vertical axis, and pro α 2(I) gene, horizontal axis, from human (Panel B). A dot is reported whenever there are three consecutive nucleotide matches.

et al. (9) and McKeon et al. (10) have examined the S1 nuclease sensitivity in the promoter region of both chicken and mouse pro α 2(I) collagen genes. These experiments have identified an S1 hypersensitive site approximately at position -200 in the chicken and -165 in the mouse promoter (9,10). In the human gene at -169 is a sequence consisting of two imperfect direct repeats closely resembling those found around the S1 hypersensitive sites in the rodent and avian genes (Fig. 1).

When the 5' region of the human pro α 2(I) gene was compared to that of the pro α 1(I) gene less than 4% homology was observed (Fig. 6B) (12). The type I genes differ also in the relative positions of the TATA and CAT boxes as well as the pyrimidine-rich segment (12). Notwithstanding these short stretches of homologies, shared by most genes, it is clear that the chromosomal area thus far analyzed in the type I genes lacks any common DNA elements, possibly responsible for collagen-specific regulation.

More striking are the similarities observed between the pro α 1(I) and pro α 2(I) genes around the initiation site of translation (Fig. 5) (12). We have already mentioned the presence of a quasi-palindrome in most of the fibrillar collagen genes, which potentially can form a hairpin structure around the start of translation (6,7). Despite variations in the length of the stems and composition of the loops, a highly conserved sequence (CATGTCCTA-n-TAGACATG) is present in relatively identical positions in the type I genes of man, mouse and chicken (6-8, 11,12). The chick pro α 1(III) gene exhibits the same consensus sequence with a single nucleotide change (7). Finally this conserved sequence is not present around the initiation site of translation of the rat pro α 1(II) gene (19). Although it has been hypothesized that these hairpin structures may play a role in the regulation of collagen biosynthesis, their functional significance has still to attain experimental confirmation (7,12).

N-propeptide domain

Biochemical and molecular studies have indicated that the N-propeptide region of the fibrillar collagen genes exhibits a high degree of structural variations among the different chains as well as in the same molecule between different species (1,13). Regardless of these differences, the N-propeptide domains of all collagen genes analyzed thus far are encoded by six exons, variable in size and composition (13,17). This observation was dramatically confirmed by the finding of two small exons coding for 9 2/3 amino acid residues of the chicken pNa2(I) (17). It has been reasoned that such small exons may represent remnants of larger units coding for the N-terminal globular region absent in the pNa2(I) (13,21). However, it is not clear what kind of evolutionary pressure could have fixed this exon-intron arrangement in the region of the gene coding for the N-terminal prepropeptide domain. We have sequenced four of the six exons coding for the human pro α 2(I) N-terminal prepropeptide, and indirectly confirmed the original observation of Tate et al. (17). Although we have thus far been unable to identify the two small exons in the 4.0 kb of genomic DNA separating exons 52 from exon 49, the ana-

lysis of the coding sequence of the former implied the existence of at least one of them.

The 205 bp of the last exon code for the 135 nucleotide long 5' untranslated region of the *pro α 2(I)* mRNA and for the first 23 1/3 amino acid residues of the protein molecule (Fig. 2,4). Twenty-one of the amino acid residues coded by exon 52 comprise the signal peptide, including the signal peptidase cleavage site (Cys+Gln) and the first 1 1/3 residues of the *pNa2(I)* (Fig. 2). The corresponding exon of the *pro α 1(I)* gene is slightly larger (222 bp) mostly because it contains a long stretch of acidic amino acid residues following the signal peptidase cleavage site (13). The N-terminal triple helical domain of the *pNa2(I)* is similarly distributed between two exons. As in the human *pro α 1(I)* gene as well as in the mouse and chick *pro α 2(I)* gene, the first exon of the N-terminal triple helical region of the human *pro α 2(I)* gene is 36 bp long and codes for four Gly-X-Y repeats (8,12,17). The remaining nine Gly-X-Y repeats of the human *pNa2(I)* are coded by the following exon (93 bp) which also contains the first four residues of the short non-helical domain of the *pNa2(I)* (Fig. 2) (1). A similar coding distribution has been observed in the other collagen genes (13,17). However, the homologous exon of the avian gene contains one less Gly-X-Y repeat resulting in a coding unit 84 bp long (17). The corresponding *pro α 1(I)* exon, 102 bp long, codes instead for three additional Gly-X-Y repeats (13).

Sequencing of the 1.6 kb Hind III genomic fragment (Fig. 1) has identified the location of the junction exon which contains the transition from the N-propeptide to the triple helical domain (Fig. 2). Thus, the entire *pNa2(I)* is contained within almost 5 kb of chromosomal DNA. This inter-dispersion value is more than double of that observed for the *pro α 1(I)* and it does reflect the average larger size of the *pro α 2(I)* introns in the entire gene (13). Comparison of the human *pNa2(I)* sequences with those derived from the avian gene showed some degree of homology (75%) in the N-terminal triple helical region, but a significant degree of variation in the N-terminal telopeptide segment coded by the junction exon (17). The dissimilarity was found clustered in the stretch immediately following the N-propeptidase cleavage site (Ala+Gln). An analogous situation has been observed in the C-terminal propeptides of the *pro α 2(I)* and *pro α 1(III)* chains, which strongly suggest a merely connective role for both of the telopeptide domains (22).

SUMMARY

In conclusion, our studies have determined for the first time the structure of the promoter region of the human *pro α 2(I)* collagen gene as well as of

most of the N-terminal prepropeptide. Comparison with the homologous region from other collagen genes has confirmed the existence of conserved sequences around the start site of translation. Comparison of the area immediately preceding the cap site in the type I genes has not identified any collagen-specific control element. The information gathered from these studies will be extremely valuable in the molecular characterization of those variants where biochemical evidence have suggested abnormalities affecting either pro α 2(I) collagen expression or normal processing of the N-terminal propeptide (5,24).

ACKNOWLEDGEMENTS

The authors are very much in debt to Drs. de Crombrughe and Schmidt for providing us with the mouse probe. We also wish to thank Ms. L. Bertha and K. Valentine for typing the manuscript. This work was supported in part by grants from the National Institute of Health (AM-32380), the March of Dimes Birth Defects Foundation and the Southern African Council for Scientific and Industrial Research. Dominique Weil is on leave of absence from Unité 12 INSERM-Hôpital des Enfants Malades, Paris, France.

REFERENCES

1. Bornstein, P. and Sage, H. (1980) Ann. Rev. Bioch. **49**:957-1003.
2. Prockop, D.J. and Kivirikko, K. (1984) N. Engl. J. Med. **311**:376-386.
3. Adams, S.L., Sobel, M.E., Howard, B.H., Olden, K., Yamada, Y., de Crombrughe, B. and Pastan, I. (1977) Proc. Natl. Acad. Sci. USA **74**:3399-3403.
4. Marsilio, E., Sobel, M.E. and Smith, B.D. (1984) J. Biol. Chem. **259**:1401-1404.
5. de Wet, W.J., Pihlajaniemi, T., Myers, J.C., Kelly, T.E. and Prockop, D.J. (1983) J. Biol. Chem. **258**:7721-7728.
6. Vogeli, G., Ohkubo, H., Sobel, M.E., Yamada, Y., Pastan, I. and de Crombrughe, B. (1981) Proc. Natl. Acad. Sci. USA **78**:5334-5338.
7. Yamada, Y., Mudryj, M. and de Crombrughe, B. (1983) J. Biol. Chem. **258**:14914-14919.
8. Schmidt, A., Yamada, Y. and de Crombrughe, B. (1984) J. Biol. Chem. **259**:7411-7415.
9. McKeon, C., Schmidt, A. and de Crombrughe, B. (1984) J. Biol. Chem. **259**:6636-6640.
10. Finer, M.H., Fodor, F.J., Boedtker, H. and Doty, P. (1984) Proc. Natl. Acad. Sci. USA **81**:1659-1663.
11. Harbers, K., Kuehn, M., Delius, H. and Jaenisch, R. (1984) Proc. Natl. Acad. Sci. USA **81**:1504-1508.
12. Chu, M.-L., de Wet, W., Bernard, M. and Ramirez, F. (1985) J. Biol. Chem. **260**:2315-2320.
13. Chu, M.-L., de Wet, W., Bernard, M., Ding, J.F., Morabito, M., Myers, J., Williams, C. and Ramirez, F. (1984) Nature **310**:337-340.
14. Maxam, A.M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA **74**:560-564.

15. Berk, A.J. and Sharp, P.A. (1978) Proc. Natl. Acad. Sci. USA 75:1274-1278.
16. Myers, J.C., Dickson, L.A., de Wet, W.J., Bernard, M.P., Chu, M.-L., Di Liberto, M., Pepe, G., Sangiorgi, F.O. and Ramirez, F. (1983) J. Biol. Chem. 258:10128-10135.
17. Banerjee, A.K. (1980) Microbiol. Review 44:175-205.
18. Kohno, K., Martin, G.R. and Yamada, Y. (1984) J. Biol. Chem. 259:13668-13673.
19. Tate, V.E, Finer, M.H., Boedtke, H. and Doty, P. (1983) Nucleic Acids Res. 11:91-103.
20. Maizel, J.V. and Lenk, R.P. (1981) Proc. Natl. Acad. Sci. USA 78:7665-7669.
21. Aho, S., Tate, V. and Boedtke, H. (1984) Nucl. Acids Res. 12:6117-6125.
22. Chu, M.-L., Weil, D., de Wet, W., Bernard, M., Sippola, M. and Ramirez, F. (1985) J. Biol. Chem. (in press).
23. Steinmann, B., Tuderman, L., Peltonen, L., Martin, G.R., McKusick, V.A. and Prockop, D.J. (1980) J. Biol. Chem. 255:8887-8893.