Volume 11 Number 16 1983 Nucleic Acids Research

Multiple 3' ends of the chicken pro α2(I) collagen gene

Sirpa Aho, Valerie Tate and Helga Boedtker

Department of Biochemistry and Molecular Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA 02138, USA

Received 25 May 1983; Revised and Accepted 25 July 1983

ABSTRACT

The precise location of the 3' ends of the chicken pro α2(I) collagen gene have been identified by S1 nuclease protection of overlapping genomic fragments by calvaria poly A containing RNA and size determination of the protected fragments on DNA sequencing gels. The gene ends 300 and 306 bp and 754 and 777 bp from the translation stop codon. The two sets of ends explain the major and minor pro α2(I) collagen mRNAs previously observed, which may result from either RNA polymerase readthrough of the first termination site and/or different processing sites.

INTRODUCTION

The structure of the chicken pro α2(I) collagen gene with its fifty odd exons distributed over 38,000 bp of genomic DNA has by now been well documented (1-4). The 5' end of this gene was precisely located by S-1 nuclease protection, primer extension (5), in vitro transcription (6) experiments and by cloning cDNA to the 5' end of the mRNA (7). The precise location of the 3' end of the gene however, was only deduced from the DNA sequence of cDNA clones: the eight terminal adenines were identified as the beginning of the poly A tail since they are preceeded by two overlapping canonical poly A addition sites, AATAAA (8). In addition, the existance of 300 bp of untranslated but transcribed genomic sequence was consistent with the size of the 3' most exon determined by electron microscopy (2,9). However, we had previously found that both the pro α1(I) and the pro α2(I) collagen cDNAs clones (10,11) hybridize to a major and at least one minor procollagen mRNA species that differ in size by 1500 and 500 bp respectively (12). The larger minor species is moreover too abundant to be an incompletely processed precursor, or result from cross hybridization to a different procollagen mRNA. Since only a single 5' end of the mRNA has been identified, it seemed likely that the two mRNAs differed only in the size of the 3' untranslated regions. To obtain a clear identification of the 3' end of the pro α2(I) gene, and to establish that the

larger mRNA resulted from either RNA polymerase readthrough of the first termination site, or from an alternative processing, the end or ends of the gene were mapped by S1 nuclease protection of genomic fragments by procollagen mRNA and by sizing the protected fragments on DNA sequencing gels.

MATERIALS AND METHODS

1. Isolation of DNA

A 1.3 kb EcoRI - Bgl II fragment located at the 5' end of the 3.7 EcoRI fragment of α2CG241, a pro α2(I) collagen gene clone (9), was subcloned in pBR322 and then separated from pBR322 fragments on a 6% acrylamide gel. Overlapping fragments were produced by cutting the fragment with Hinf I and Dde I (New England Biolabs). The 3' ends were labelled using α-P^{32} labelled nucleotide triphosphates and the large fragment of E. coli DNA polymerase I (New England Biolabs) (13). The labelled fragments were isolated from an 8% acrylamide, 7% urea denaturing gel. The strands were separated by denaturing the fragment with 30% Me_2SO and then electrophoresing on a 5% acrylamide, 0.1% bisacrylamide gel (13). Separated complementary strands were used for S1-nuclease protection experiments.

2. DNA sequence determinations

The nucleotide sequence of the separated strands was determined as described by Maxam and Gilbert (13).

3. RNA isolation

Embryonic chick calvaria poly A containing RNA was isolated as described by Tate et al. (7).

4. S1 nuclease protection studies

The isolated strands from Hinf I and Dde I fragments were hybridized to calvaria poly A containing RNA and then treated with S1 nuclease as described by Favolaro et al. (14), except that formamide was omitted from the hybridization. In brief, a twofold molar excess (≈20,000 cpm) relative to the pro α2(I) collagen mRNA content of calvaria poly A containing RNA (0.05 to 0.1μg) in 40 mM Pipes pH 6.4, 1 mM EDTA, 0.4 M NaCl, were sealed in glass capillary tubes, heated 2' at 90° C and then incubated at 45° C for 24 h. After incubation the capillaries were opened and the 10 μl samples were diluted into 400 μl of icecold S1 buffer, 30 mM Na Acetate pH 4.6, 4 mM $ZnSO_4$, 0.28 M NaCl, 5% glycerol, containing 6.25 μg/ml of salmon sperm DNA. 1 μg (113 U) of S1 nuclease (Bethesda Research Laboratories) was added and reaction was carried out at 15° C for 2 h. The reaction was stopped by addition of 80 μl

0.5 M Tris-HCl pH 9.5, 1 M NaAc, 20 mM EDTA and two volumes of ethanol. After precipitation and washing, the samples were run on a denaturing 8% acrylamide, 7% urea gel. As controls, identical hybridizations were carried out but either no S-1 nuclease was added, the RNA was omitted, or calvaria poly A containing RNA was replaced by 0.1 µg of total yeast RNA.

5. All experiments were carried out under the guidelines established by the N.I.H.

RESULTS

To locate the 3' end or ends of the pro α2(I) collagen gene, adjacent and overlapping Hinf I and Dde I restriction fragments were identified and isolated from a 1300 bp EcoRI - BglII fragment within which the gene ends. The arrangement of the fragments is shown in Fig. 1. The 593 bp and 352 bp Hinf I fragments and the 417 bp and 357 bp Dde I fragments were 3' end labelled, denatured, strand separated, and hybridized to calvaria poly A^+ RNA, as described in MATERIALS AND METHODS. The RNA protected two fragments about 200 bp long as well as the entire 593 bp Hinf I fragment when the latter was hybridized to it, as shown in lane B2 in Figure 2. This strongly suggests that one pro α2(I) mRNA ends 200 bp 3' to the 5' most Hinf I site while the other extends 3' to the second Hinf I site. This suggestion was confirmed when the 352 bp Hinf I fragment was hybridized to calvaria poly A^+ RNA and

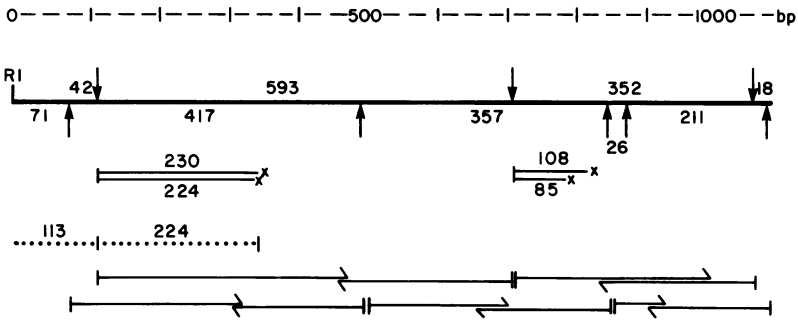

Figure 1. Restriction Map of the 3' ends and flanking regions of the chick pro α2(I) collagen gene. Distances are shown in base pairs (bp). ↓, arrows pointing down are Hinf 1 sites, ↑, arrows pointing up Dde I sites. The solid lines ending in X show the size of the two protected fragments obtained when the 593 bp and the 352 bp Hinf 1 fragment were hybridized to calvaria mRNA prior to S1 nuclease digestion. The dotted line represents the part of the fragment whose sequence was reported previously (8). Horizontal arrows at the bottom show fragments whose DNA sequence has been determined.

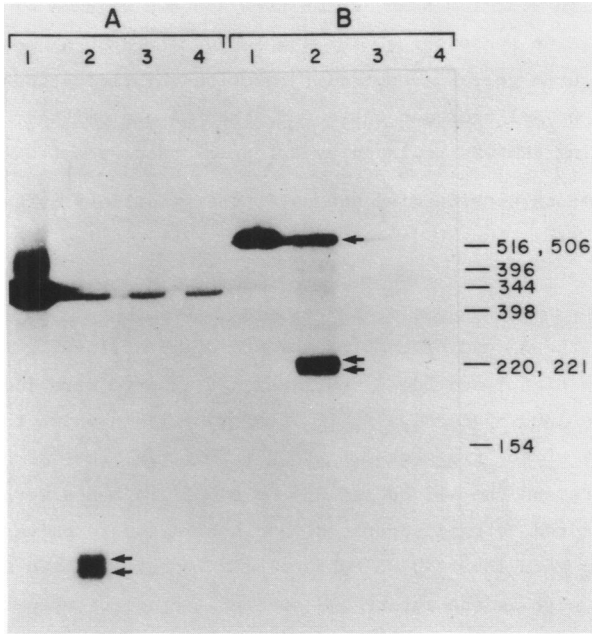

Figure 2. Protection of genomic fragments by calvaria mRNA from digestion with S1 nuclease. A. Hybridization of complementary strand of 352 bp Hinf 1 fragment to calvaria RNA without S1 nuclease digestion (lane 1), with nuclease digestion (lane 2), without RNA (lane 3) and with yeast RNA (lane 4). The 352 bp fragment that is protected from S1 digestion in the absence of RNA, or by yeast RNA must be a double stranded DNA contaminant of the isolated strand. B. Same as A, except 593 bp Hinf 1 fragment was used.

only two fragments, about 100 bp long, were protected, as shown in Fig. 2, lane A2. Lanes 1, 3, 4 in Fig. 2 are the controls: minus S1 nuclease, minus RNA, hybridized to yeast RNA. The four protected fragments which mark the four ends of the gene are identified in the map in Fig. 1. Their precise size was determined on a DNA sequencing gel (data not shown).

The DNA sequence of the first 335 bp at the 5' end of the RI-Bgl II fragment had been determined by sequencing pro α2(I) collagen cDNA clones (8). In order to verify the cDNA sequence as well as to examine the DNA sequence surrounding the ends of the gene, the DNA sequence of some 1000 bp was determined following the sequencing scheme shown at the bottom of Fig. 1. The results presented in Fig. 3 show four canonical poly A addition sites, identified in boxes numbered 1 through 4. These are located 21 to 22 bp from the end of the gene, identified by one or more arrows over the terminal

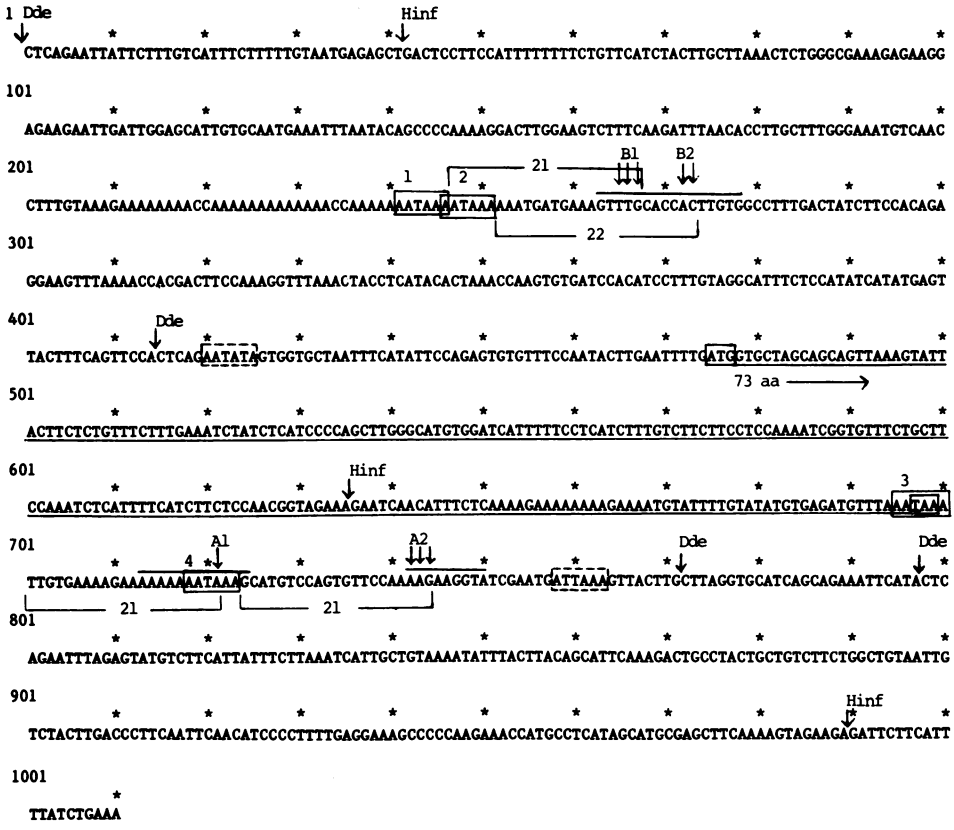

Figure 3. DNA sequence of 1 kb of 3' untranslated and flanking gene regions of the pro α2(I) collagen gene. Four canonical poly A addition sites are in boxes labelled 1,2,3, and 4. Two related poly A addition sites, used in other genes, are each enclosed by a dashed box. The four most prominent protected fragments end at several adjacent nucleotides identified by arrows at B1, B2, A1 and A2. The longest open reading frame is underlined, with the ATG and TAA start and stop signals enclosed in boxes.

nucleotides. These 3' untranslated and 3' flanking gene regions are very AT rich with a GC content of only 35%. This is in sharp contrast to the 56% GC content of 5' untranslated region and the 67.5 GC content of the 400 bp 5' flanking gene region (4). The high AT content is marked by several runs of A's which of course provide possible poly A addition sites which differ from the canonical sequence by one nucleotide. Other sequences found in this region which differ by only one nucleotide include ATTAAA, at 758, the poly A addition site used in the chicken lysozyme gene (15) and the mouse pancrease α-amylase gene (16) and AATATA at 420, also used in the mouse α-amylase gene

(17). Other than the poly A addition site, the nucleotides at the four ends of the gene have little or no homology. The 1000 bp sequence in Fig. 3 has several open reading frames including one located between the two sets of ends of the gene. The latter codes for 73 amino acids, many of which are hydrophobic. The sequence begins with AUG in position 475 and ends with UAA in position 697 and hence may be translated *in vivo*. Unlike the hydrophobic exons in the same region of the mouse IgM gene (18), there is no evidence that this 219 bp region is ever spliced to a donor site 5' to the termination codon. Finally the first two ends of the gene, B1 and B2, are some 450 and 475 bp from the second two ends, A1 and A2, as predicted by the 500 bp difference in the size of the major and minor pro α2(I) collagen mRNAs (12).

DISCUSSION

The precise location of the 3' end of a gene, arbitrarily defined as the nucleotide to which adenines are added post transcriptionally, rarely coincides with the 3' end of the primary transcripts of the gene, in the relatively few instances where these have been determined (19,20). We do not know if the four poly A addition sites identified at the 3' end of the pro α2(I) collagen gene are also RNA polymerase termination sites, or serve as processing signals. The latter seems the more likely possibility since RNA polymerase is clearly capable of reading through the first set of overlapping signals and proceeding an additional 450 to 475 bp downstream.

Chicken type I procollagen genes are not the only collagen genes which have multiple transcripts for the same single copy gene. The chicken α1(II) procollagen gene also has at least two transcripts (21) while the human α2(I) gene has four transcripts (22). Only for the chicken type II collagen gene is there evidence for differential expression of the two transcripts during development (21). Multiple transcripts of different sizes from single copy genes have also been reported for the mouse α-amylase gene (16,17), the dihydrofolate reductase gene (23), the chicken X gene (24), the mouse immunoglobin IgM gene (18), and quite recently, the *Drosophila melanogaster* myosin heavy chain gene (25), the chicken vimentin gene (26), and the β-microglobulin gene (27). These can result from different splicing modes at 5' end of the gene, as is the case of the α-amylase (16), or at the 3' end, as is the case of the *Drosophila melanogaster* myosin heavy chain (25) and the mouse IgM (20) genes with different modes used in different tissues or at different times during development. Alternatively they can reflect RNA polymerase read-through at the 3' end of the gene and alternate poly A recognition and

addition signals whose use is not developmentally regulated as appears to be the case in the chicken vimentin (26) and pro α2(I) collagen genes. Since the AATAAA sequence has been highly conserved, and since processing and/or poly adenylation absolutely requires the preservation of this sequence (20), it may well have proved beneficial to incorporate more than one copy of this sequence to insure the correct generation of a polyadenylated 3' end.

The AAUAAA sequence is clearly necessary but not sufficient for processing and/or polyadenylation to proceed efficiently since this sequence occurs frequently in collagen introns. Earlier suggestions that a specific hairpin secondary structure may be involved (28) became unlikely after many 3' ends of poly A containing RNAs had been sequenced which could not form such a structure (19,29). The DNA sequences and factors required for efficient processing at the 3' end have received relatively little attention, especially when compared to the many studies of the sequences required for efficient and precise initiation of primary transcripts. However, an understanding of efficient termination or processing of collagen genes is a challenge. One cannot help but wonder what sort of sequence or structure the RNA polymerase recognizes as signalling termination after it has transcribed some 38,000 bp, of which 34,000 bp are AT rich intron sequences which contain several copies of the canonical AATAAA sequence that do not function as poly A addition sites.

ACKNOWLEDGEMENTS

We want to thank Elizabeth Levine for technical assistance, Mitchell Finer and Gary Brennan for help with computer programs for handling DNA sequence data, Nancy Pegg for typing this manuscript, and the Academy of Science of Finland for support to Sirpa Aho. This research was supported in part by a grant from the N.I.H.

REFERENCES

1. Vogeli, G., Ohkubo, H., Avvedimento, V.E., Sullivan, M., Yamada, Y., Mudryj, M., Pastan, I. and de Crombrugghe, B. (1980) Cold Spring Harbor Laboratory Symposium 45, 777-783.
2. Ohkubo, H., Vogeli, G., Mudryj, M., Avvedimento, V.E., Sullivan, M., Pastan, I and de Crombrugghe, B. (1980) Proc. Natl. Acad. Sci. USA 77 7059-7063.
3. Wozney, J., Hanahan, D., Tate, V., Boedtker, H., and Doty, P. (1981) Nature 294, 129-135.
4. Tate, V., Finer, M., Boedtker, H. and Doty, P. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 1039-1049.
5. Vogeli, G., Ohkubo, H., Sobel, M.E., Yamada, Y., Pastan, I and de Crombrugghe, B. (1981) Proc. Natl. Acad. Sci. USA 78, 5334-5338.
6. Merlino, G.T., Vogeli, G., Yammamato, T., de Crombrugghe, B. and Pastan,

I. (1981) J. Biol. Chem. 256, 11251-11258.
7. Tate, V., Finer, M., Boedtker, H. and Doty, P. (1983) Nucleic Acids Res. 11, 91-104.
8. Fuller, F. and Boedtker, H. (1981) Biochemistry 20, 996-1006.
9. Wozney, J., Hanahan, D., Morimoto, R., Boedtker, H., and Doty, P. (1981). Proc. Natl. Acad. Sci. USA 78, 712-716.
10. Lehrach, H., Frischauf, A.M., Hanahan, D., Wozney, J., Fuller, F., and Boedtker, H. (1979) Biochemistry 18, 3146-3152.
11. Lehrach, H., Frischauf, A.M., Hanahan, D., Wozney, J., Fuller, F., Crkvenjakov, R. Boedtker, H. and Doty, P. (1978) Proc. Natl. Acad. Sci. USA 75, 5417-5421.
12. Rave, N., Crkvenjakov, R., and Boedtker, H. (1979) Nucleic Acids Res. 11 3559-3567.
13. Maxam, A. and Gilbert, W. (1980) In 'Methods in Enzymology' (L. Grossman and K. Moldave, eds.) Vol. 65, pp. 499-560. Academic Press, N.Y. and London.
14. Favaloro, J., Treisman, R., and Kamen, R. In 'Methods in Enzymology' (L. Grossman and K. Moldave, eds.) Vol. 65, pp. 718-735. Academic Press, N.Y. and London.
15. Jung, A., Sippel, A.E., Grez, M., Schutz, G. (1980) Proc. Natl. Acad. Sci. USA 77, 5759-5763.
16. Hagenbuchle, O, Bovey, R., Young, R.A. (1980) Cell 21, 179-187.
17. Tosi, M., Young, R.A., Hagenbuchle, O. and Schibler, U. (1981) Nucleic Acids Res. 9, 2313-2322.
18. Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L. and Wall, R. (1980) Cell 20, 303-312.
19. Breathnach, R. and Chambon, P. (1981) Ann. Rev. Biochem. 50, 349-383.
20. Fitzgerald, M. and Shenk, T. (1981) Cell 24, 251-260.
21. Vuorio, E., Sandell, L., Kravis, D. Sheffield, V.C., Vuorio, T., Dorfman, A. and Upholt, W.B. (1982) Nucleic Acids Res. 10, 1175-1192.
22. Myers, J.C., Dickson, L.A. deWet, W.J., Bernard, M.P., Chu, M., DiLiberto, M., Pepe, G., Sangiargi, F.O. and Famiree, F. (1983) J. Biol. Chem. in press.
23. Setzer, D.R., McGrogan, M., Nunberg, J.H. and Schimke, R.T. (1980) Cell 22, 361-370.
24. Heilig, R., Perrin, F., Gannon, F., Mandel, J.L. and Chambon, P, (1980) Cell 20, 625-637.
25. Rozek, C.E. and Davidson, N. (1983) Cell 32, 23-34.
26. Zehner, Z.E. and Paterson, B.M. (1983) Proc. Natl. Acad. Sci. USA 80, 911-915.
27. Parnes, J.R., Robinson, R.R. and Seidman, J.G. (1983) Nature 302, 449-452.
28. Proudfoot, N.J., and Brownlee, G.G. (1974) Nature 252, 359-362.
29. Proudfoot, N.J., Cheng, C.C. and Brownlee, G.G. (1976) Prog. in Nucleic Acid Res. and Mol. Biol. (W.E. Cohn and E. Bolkin, eds.) Vol. 19, pp. 123-133. Academic Press, N.Y. and London.