recognition domain and the availability of this domain are essential for the staphylococcal clumping reaction.

References

Chen, R., & Doolittle, R. F. (1970) *Biochemistry 10*, 4486.

Cierniewski, C. S., Kloczewiak, M., & Budzynski, A. Z. (1986) *J. Biol. Chem. 261*, 9116.

Dang, C. V., Ebert, R. F., & Bell, W. R. (1985) *J. Biol. Chem. 260*, 9713.

Dejana, E., Languino, L. R., Polentarutti, N., Balconi, G., Ryckewaert, J. J., Larrieu, M. J., Donati, M. B., Mantovani, A., & Marguerie, G. (1985) *J. Clin. Invest. 75*, 11.

Durack, D. T. (1975) *J. Pathol. 115*, 81.

Fling, S. P., & Gregerson, D. S. (1986) *Anal. Biochem. 155*, 83.

Furlan, M., & Beck, E. A. (1975) *Thromb. Res. 7*, 827.

Gonda, S. R., & Shainoff, J. R. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 4565.

Halsall, M. B. (1967) *Nature (London) 215*, 880.

Haverkate, F., & Timan, G. (1977) *Thromb. Res. 10*, 803.

Hawiger, J., Timmons, S., Strong, D. D., Cottrell, B. A., Riley, M., & Doolittle, R. F. (1982a) *Biochemistry 21*, 1407.

Hawiger, J., Timmons, S., Kloczewiak, M., Strong, D. D., & Doolittle, R. F. (1982b) *Proc. Natl. Acad. Sci. U.S.A. 79*, 2068.

Kloczewiak, M., Timmons, S., Lukas, T. J., & Hawiger, J. (1984) *Biochemistry 23*, 1767.

McNamana, G., Lindon, J. N., Kloczewiak, M., Smith, M. A., Hawiger, J., & Salzman, E. W. (1986) *Blood 68*, 363.

Means, G. E., & Feeney, R. E. (1968) *Biochemistry 7*, 2192.

Strong, D. D., Laudano, A., Hawiger, J., & Doolittle, R. F. (1982) *Biochemistry 21*, 1414.

Timmons, S., & Hawiger, J. (1984) *Circulation 70*, II-96.

Timmons, S., & Hawiger, J. (1986) *Trans. Assoc. Am. Physicians 99*, 226–235.

Timmons, S., Kloczewiak, M., & Hawiger, J. (1984) *Clin. Res. 31*, 498A.

Van der Drift, A. C. M., & Poppema, A. (1982) in *Fibrinogen: Recent Biochemical and Medical Aspects* (Henschen, A., Graeff, H., & Lottspeich, F., Eds.) p 47, de Gruyter, Berlin.

A Gene for Rabbit Synovial Cell Collagenase: Member of a Family of Metalloproteinases That Degrade the Connective Tissue Matrix†

M. Elizabeth Fini,‡ Izabela M. Plucinska,‡ Anyce S. Mayer,‡ Robert H. Gross,§ and Constance E. Brinckerhoff*,‡,‖

Departments of Medicine and Biochemistry, Dartmouth Medical School, and Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire 03756

Received March 9, 1987; Revised Manuscript Received May 15, 1987

ABSTRACT: We have determined the nucleotide sequence of a collagenase mRNA from rabbit synovial cells from which the primary structure of the encoded protein was deduced. This proteinase is 51% homologous to the enzyme that activates it from the zymogen form, rabbit synovial cell activator/stromelysin. Rabbit collagenase and activator/stromelysin thus share comembership in a gene family that includes human skin collagenase; the human and rabbit metalloproteinase, activator/stomelysin; and an oncogene-induced proteinase from rat named transin. The mRNA sequence of collagenase enabled us to completely map the structure of its gene, which is 9.1 kilobases and is composed of 10 exons and 9 introns. This is the first report of the structure of a collagenase gene. We show that it has striking similarity to additional members of this metalloproteinase gene family, transin genes I and II of rat. We have further sequenced genomic DNA flanking the collagenase gene and have identified nucleic acid elements of possible importance in gene regulation.

Collagen is the most prevalent protein in the body, and its metabolism plays a central role in many normal biological processes (Gross, 1982; Hay, 1984). The capacity to control collagen degradation also appears to be an essential characteristic of malignant tumors, allowing their successful invasion of surrounding tissues (Liotta et al., 1984; Mignatti et al., 1986). Furthermore, a number of other pathological conditions are marked by inappropriate or excessive collagenolysis (Wooley & Evanson, 1980). The metalloproteinase collagenase is the rate-limiting enzyme in this process. It is secreted from cells as an inactive proenzyme and can be activated by other proteinases found in the extracellular matrix (Harris et al., 1984). In rabbits, conversion of procollagenase to the active enzyme appears to require a specific metalloproteinase called activator, which is itself secreted as a proenzyme (Vater et al., 1983). Thus the control of collagenolysis in rabbits appears to be regulated via a minicascade of proteinases. Recently, rabbit activator has been shown to be identical with stromelysin, a metalloproteinase originally isolated from rabbit fibroblasts that has the ability to degrade noncollagenous matrix (Chin et al., 1985; Whitham et al., 1986; Fini et al., 1987).

This laboratory has focused on collagen turnover in rheumatoid arthritis, a disease in which excessive collagenase secretion by synovial fibroblasts lining the joints results in ex-

† Supported in part by grants from the USPHS (NIH-AR26599), from the New Hampshire and National Chapters of the Arthritis Foundation, from the RGK Foundation (Austin, TX), and from the Hitchcock Foundation.

* Correspondence should be addresed to this author.

‡ Department of Medicine, Dartmouth Medical School.

§ Department of Biological Sciences, Dartmouth College.

‖ Department of Biochemistry, Dartmouth Medical School.

0006-2960/87/0426-6156$01.50/0

tensive destruction of cartilage and bone (Harris, 1985). Successful control of joint damage in this disease may be facilitated through identification of the control points in collagenase gene expression and mechanisms regulating enzyme activity. As a model system, we have used primary cultures of rabbit synovial cells, in which collagenase synthesis can be induced in response to agents such as phorbol esters (Brinckerhoff et al., 1979) or repressed with dexamethasone or retinoids (Brinckerhoff et al., 1980, 1981; Brinckerhoff & Harris, 1981). Two species of rabbit synovial cell procollagenase can be resolved on sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) with apparent molecular masses of 57 and 61 kilodaltons (kDa) (Nagase et al., 1983). The larger form is glycosylated and can be chased into the 57-kDa protein by treatment of cell cultures with tunicamycin. Both forms can be activated to degrade collagen. Induction of collagenase, in the rabbit synovial cell system, is correlated with an increase in the mRNA levels for the enzyme (Gross et al., 1984) which is regulated, at least in part, by an increase in the message half-life (Brinckerhoff et al., 1986). Amplification of the collagenase gene is not involved in the increase in message levels (Fini et al., 1986a,b).

Our further studies on the regulation of collagenase have been hampered by a lack of structural knowledge of the protein, its mRNA, and its gene. Here we have determined the sequence of a collagenase mRNA from rabbit synovial cells. From this sequence we have deduced the primary structure of rabbit synovial cell collagenase. Finally, we have mapped the intron–exon structure of the collagenase gene from which this message was transcribed, and we have examined the DNA sequences flanking the 5′ end of the gene.

Materials and Methods

Construction and Screening of the Rabbit cDNA Library and Restriction Mapping of Clones. The construction in bacteriophage λ gt-11 of the cDNA library used to isolate collagenase clones has been previously described (Brinckerhoff et al., 1987a; Fini et al., 1987). Collagenase clone p206 (Fini et al., 1986a), which is a subclone of the 2.4-kilobase (kb) *Eco*RI fragment from collagenase genomic clone λ 13B, was used as a probe to isolate cDNA clone λ col-99. λ col-H22C was isolated by using genomic subclone p208 (Fini et al., 1986a) containing the 0.4k-kb *Eco*RI fragment from λ 13B. Screening of plaques, preparation of phage DNA, and restriction mapping were all performed according to standard methodology (Maniatis et al., 1982).

Isolation of Genomic Clones for Collagenase. The isolation of four genomic clones, λ 22A, 14B, 13A, and 13B, was described previously (Fini et al., 1986b). To isolate additional clones, the 1.5-kb *Sst*I fragment located at the 5′ end of genomic clone λ 13B (Fini et al., 1986b) was used to screen a rabbit genomic library kindly provided by Ross Hardison. This fragment contained no highly repetitive genomic sequences as determined by an assay described previously (Fini et al., 1986b) and was used to screen a trial set of 30 000 clones to be sure it did not contain low-copy repetitive sequences, before using it to screen another 600 000 plaques. Five positive clones were isolated, and one of these, λ WLT2, containing the most new chromosomal DNA, was characterized further by restriction mapping.

DNA Sequencing. *Eco*RI fragments of insert DNA from cDNA or genomic clones were cloned into an M13 phage vector and sequenced by the dideoxynucleotide termination method (Sanger et al., 1977). To sequence inserts longer than 400–500 bases, a set of ordered deletions was prepared from each original M13 recombinant by using the Cyclone kit manufactured by International Biotechnology Laboratories, New Haven, CT (Dale et al., 1985). Also, in one case, a 20-base oligonucleotide (5′-TGGACTTCAAGCTGCTTATG-3′, synthesized by the Molecular Genetics Center, Dartmouth College, Hanover, NH) was used to determine the sequence within a region of cDNA that could not be deletion subcloned. The sequence of cDNA clones was determined from both strands except from bases 934–1006 of the collagenase mRNA sequence. Genomic DNAs were sequenced only from one strand. Sequences were analyzed by using a program for the Apple Macintosh called The DNA Inspector II (Gross, 1986).

S1 Analysis. (*1*) *Mapping of the Sizes of Exons 7 and 8.* Single-stranded M13 subclones containing *Eco*RI inserts representing portions of the template strand of the collagenase gene were each hybridized to the sense strand of the 0.9-kb *Eco*RI fragment of col-99, also subcloned into M13. Annealing was in 10 μL of hybridization buffer [90 mM NaCl, 7.5 mM tris(hydroxymethyl)aminomethane (Tris), pH 8.0, 10 mM $MgCl_2$, 0.75 mM ethylenediaminetetraacetic acid (EDTA)] for 1 h at 67 °C with 1 μg of each cloned DNA. After hybridization, the reaction was diluted to 50 μL with S1 nuclease buffer, and 0.066 unit of S1 nuclease (Bethesda Research Labs, Bethesda, MD) was added. Incubation was for 1 h at 37 °C to digest single-stranded DNA. Double-stranded DNA fragments protected from S1 nuclease were sized by electrophoresis on a polyacrylamide gel. Undigested DNA was also examined to control for contamination with defective M13 miniphage DNA that could be mistaken for S1 digestion products.

(*2*) *Mapping the 5′ End of the Collagenase Gene.* S1 mapping was performed as described by Favaloro et al. (1980). The replicative form of a subclone in M13mp19 of the more 5′ of the 2.4-kb fragments of λ 13A (Figure 1) was digested with *Nco*I. This liberated a genomic fragment of 0.42 kb with its 3′ end within the middle of exon 1. The fragment was radiolabeled with polynucleotide kinase. Hybridization of 50 μg of total cell RNA from phorbol myristate acetate (PMA) induced synovial cells was performed with 50 ng of labeled DNA at 42 °C for 3 h in 80% formamide. S1-protected fragments were sized in a 6% polyacrylamide/7 M urea gel with a sequencing ladder as a size standard.

Primer Extension. A 20-base oligonucleotide with the sequence 5′-ACCCGAGGGTACCGAAGGGT-3′, complementary to bases 109–121 of the collagenase mRNA sequence, was used as a primer for cDNA synthesis from collagenase mRNA (primer prepared by the Molecular Genetics Center of Dartmouth College). One hundred nanograms of primer was annealed to 30 μg of whole-cell RNA for 5 h at 50 °C in 10 μL of buffer containing 100 mM NaCl. The reaction was cooled, and then 5 μL of [^{35}S]dATP (800 Ci/mmol) and 5 μL of RT buffer [160 mM Tris, pH 7.9, 20 mM $MgCl_2$, 8 mM dithiothreitol (DTT), 800 μM each of dCTP, dGTP, and TTP] were added. M-MLV reverse transcriptase (Bethesda Research Labs, Bethesda, MD) was mixed with RT buffer at 40 units/μL just before use. The reaction was incubated at 37 °C for 30 min and then chased with the addition of 1 μL of 10 mM dATP and 1 μL of reverse transcriptase (200 units/μL) for an additional 30 min. RNA was then hydrolyzed by making the reaction 0.5 M in NaOH for an additional 30 min at 37 °C. Finally, neutralization was performed with HCl prior to percipitation of nucleic acids with ethanol. Reaction products were analyzed on a 6% polyacrylamide/7 M urea sequencing gel with a DNA sequencing ladder as a size standard.

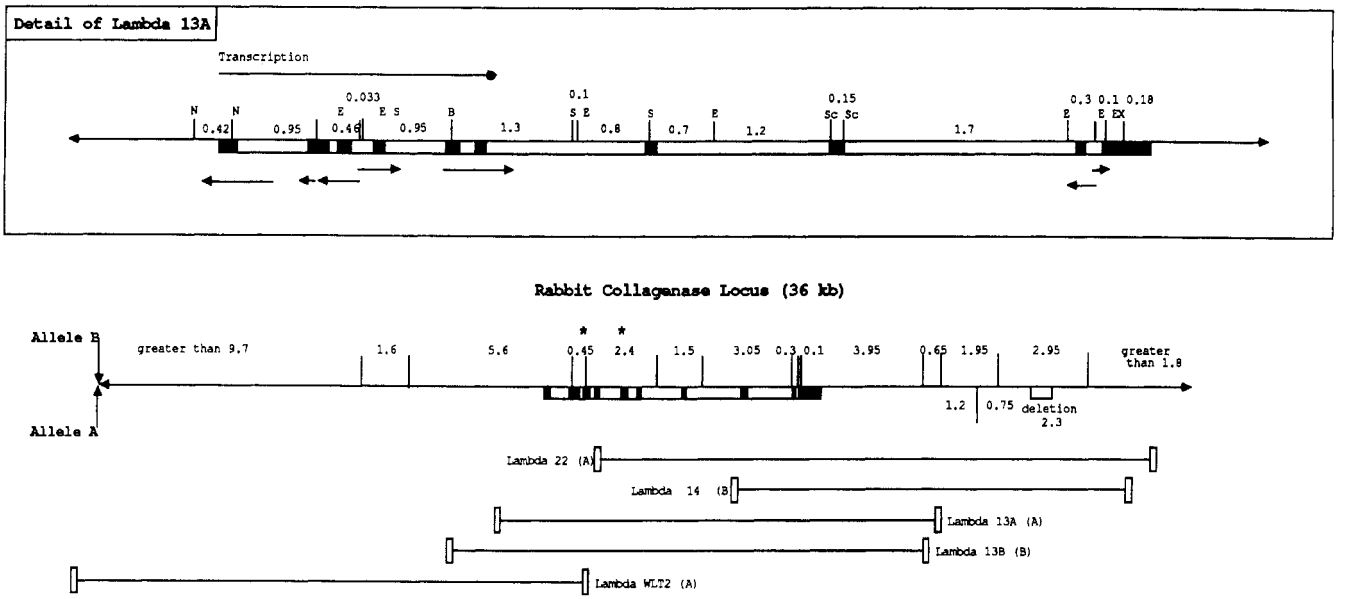

FIGURE 1: Structure of the rabbit genome at a synovial cell collagenase locus. An *Eco*RI restriction enzyme map covering 36 kb of the rabbit genome is shown. *Eco*RI sites (vertical lines) drawn above the line are those found in the B allelic version of this genetic locus. The site drawn below the line and the deletion indicated are additional features found in the A allele. Starred are the two *Eco*RI fragments subcloned into p206 (2.4 kb) and p208 (0.45 kb) and used to screen the cDNA library. The position of the collagenase gene is drawn to scale on the map with exons represented by filled boxes and introns by open boxes. Drawn below the genomic map are the overlapping clones that were isolated and characterized to define this region. The letter after the name of each clone designates it as the A or B allele. An expanded map of λ 13A, showing the collagenase gene in more detail, is boxed at the top of the figure. Arrows under the line indicate the genomic regions sequenced and the direction of sequencing. The arrow over the line shows the direction of transcription of the gene. Locations of selected restriction enzyme sites that were essential to mapping the gene are indicated. (The *Sca*I, *Nco*I, and *Bst*XI sites were mapped only in the necessary regions of the clone, and others may exist.) Symbols: E = *Eco*RI, N = *Nco*I, X = *Xba*I, S = *Sst*I, Sc = *Sca*I, B = *Bst*XI.

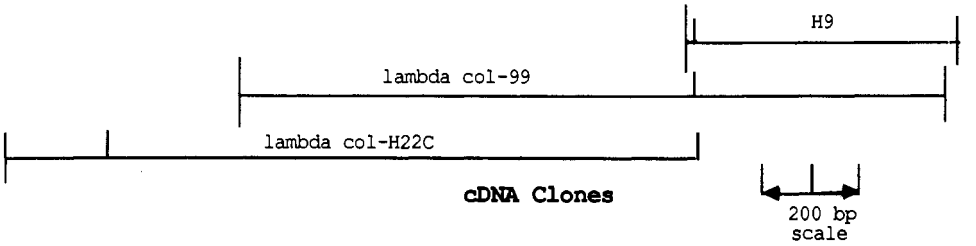

FIGURE 2: Structure of the collagenase cDNA clones. The overlapping inserts of the collagenase cDNA clones, H9, col-99, and col-H22C, are shown. *Eco*RI sites are indicated by lines and the boundaries of the cDNA by boxes. H9 is cloned into pBR322 at the *Pst*I site via homopolymer tails. col-99 and col-H22C are cloned into the *Eco*RI site of λ gt-11 via *Eco*RI linkers. The *Eco*RI site at the 3′ end of col-H22C, however, represents an internal cDNA site that was left unprotected by *Eco*RI methylase during construction of the cDNA library. Thus there is no linker on this end.

Cell Culture and Preparation of RNA. Cultures of rabbit synovial fibroblasts were established in monolayer as described (Dayer et al., 1976). For experiments, cells were plated in 100- or 150-mm diameter dishes and grown to confluence in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% horse serum and 5% fetal calf serum (Gibco, Grand Island, NY). To induce collagenase synthesis, cells were placed in serum-free medium plus 0.2% lactalbumin hydrolysate and treated with 10^{-8} M phorbol myristate acetate (Consolidated Midland, Brewster, NY). After 48 h, culture medium was tested for collagenase activity by a fibril assay (Harris et al., 1969), and RNA was prepared (Chirgwin et al., 1979; Ullrich et al., 1977) from cells that were positive for this assay.

RESULTS

Genomic Clone Structure. We have isolated and restriction mapped five different rabbit genomic clones that span 36 kb of the rabbit genome (Figure 1). Preliminary characterization of four of these clones, isolated from a rabbit genomic library with a 530 base pair (bp) cDNA clone for collagenase, suggests that they can be classified into two types, depending on the presence or absence of certain restriction sites or of a deletion 3′ to the collagenase gene (Fini et al., 1986b). While these clones could represent two different collagenase genes, we favor the hypothesis that they are allelic versions of the same gene since their sequences are strongly homologous, as judged by hybridization experiments, over their entire length. This would be unlikely if the clones represented two different genes embedded within the DNA of two different chromosomal loci. We will subsequently refer to the different genes represented in these two clone types as allele A and allele B, as they are delineated in Figure 1.

Screening for a fifth clone, λ WLT2, was performed by using a DNA fragment representing the most 5′ portion of these clones. λ WLT2 includes an additional 11.9 kb of chromosomal DNA and is important because it may contain sequences upstream from the collagenase gene that are important in its regulation.

Structure of Collagenase cDNA Clones. The structure of two newly constructed cDNA clones for collagenase, λ col-99 and λ col-H22C, is diagrammed in Figure 2. These clones overlap and considerably extend the previously described 530-bp cDNA clone H9, which represents the 3′ end of the collagenase mRNA (Gross et al., 1984; Fini et al., 1986a,b). The complete DNA sequence of col-99 was determined, as was the region of col-H22C that represented the nonoverlapping

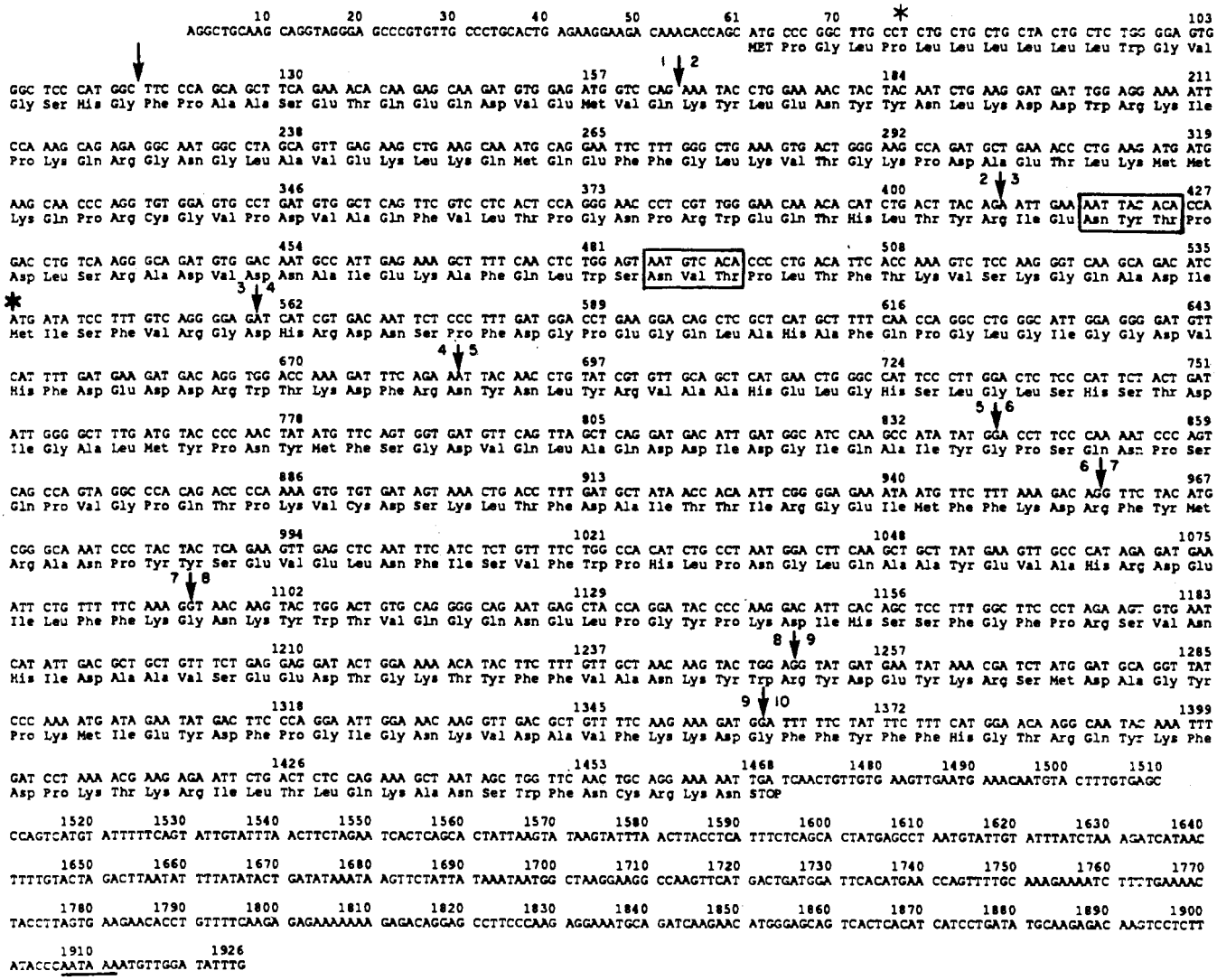

FIGURE 3: Sequence of the synovial cell collagenase cDNA. The sequence of the synovial cell collagenase cDNA (allele type A) derived from the cDNA clones, λ col-99 and λ col-H22C, as well as from the genomic clone, λ 13A, is shown. The 5′ ends of λ col-H22C and col-99 are starred. The primary structure of the predicted protein is indicated under the sequence. An arrow marks the proposed signal peptide cleavage site, and boxes are placed around the two potential sites for N-linked glycosylation. Arrows point to the splice junctions, and the exon from which the sequence block was derived is indicated. (Note that the boundary between exons 7 and 8 is only an approximation as discussed in the text.) Finally, the consensus sequence for poly(A) addition is underlined.

portion of the mRNA. However, to ensure that col-99 and col-H22C were derived from the same mRNA, approximately 150 bases on the 3′ end of col-H22C was sequenced. The sequence of the new clones (Figure 3) was compared with that of H9. Of the sequences shared by col-99 and col-H22C and by H9, all were identical, except for three single base changes located at sequence positions 1546 (T substituted for C), 1739 (A substituted for G), and 1867 (A substituted for G). Since all differences found are in the 3′ untranslated regions of the cDNA, they do not affect the amino acid sequence. The base change at position 1546 in the new clones creates an *Xba*I site and establishes col-99 and col-H22C as members of the allele A group (Fini et al., 1986b), while H9, which does not contain this *Xba*I site, (Fini et al., 1986a), has been classified as a member of the allele B class.

5′ End of the Collagenase Gene and mRNA. The collagenase message in rabbit synovial fibroblasts has been sized at 2.2 kb (Fini et al., 1986a). Since the full length of the cDNA covered by all three clones shown in Figure 2 is only about 1.9 kb, we suspected that we had not achieved a full-length copy of the message. The length of the remaining distance to the 5′ end of the collagenase message was determined by primer extension of a 20-base oligonucleotide complementary to collagenase mRNA and corresponding to bases 27–46 of λ col-H22C. [The position of the primer-complementary sequence on the collagenase gene is shown in Figure 7 (see below).] The resulting extension product, sized on a polyacrylamide gel, was a predominant band of 121 nucleotides in length (Figure 4A). This indicates that λ col-H22C is 75 bases short of full length on the 5′ end.

Since the sequence of the 5′ end of the collagenase message was not cloned in cDNA, we determined its sequence from the genomic clone, λ 13A. Sequencing was performed in the region homologous to the 5′ end of λ col-H22C and 348 bases further upstream. Then, S1 nuclease mapping was used to establish the 5′ boundary of the collagenase gene exon in this region and thus define those regions of sequence that would be represented in mRNA. The probe used for this analysis was an *Nco*I fragment, located within the 2.4-kb *Eco*RI fragment at the 5′ end of λ 13A (Figure 1). The *Nco*I cleavage site at the 3′ end of this fragment corresponds to base 34 of cDNA clone λ col-H22C on the sense strand (See Figure 7 below). This is base 38 of the antisense strand which hybridizes to mRNA in the S1 analysis. The 5′ end lies within a nontranscribed area of DNA. This fragment was 5′ end labeled, hybridized to RNA from PMA-induced synovial cells,

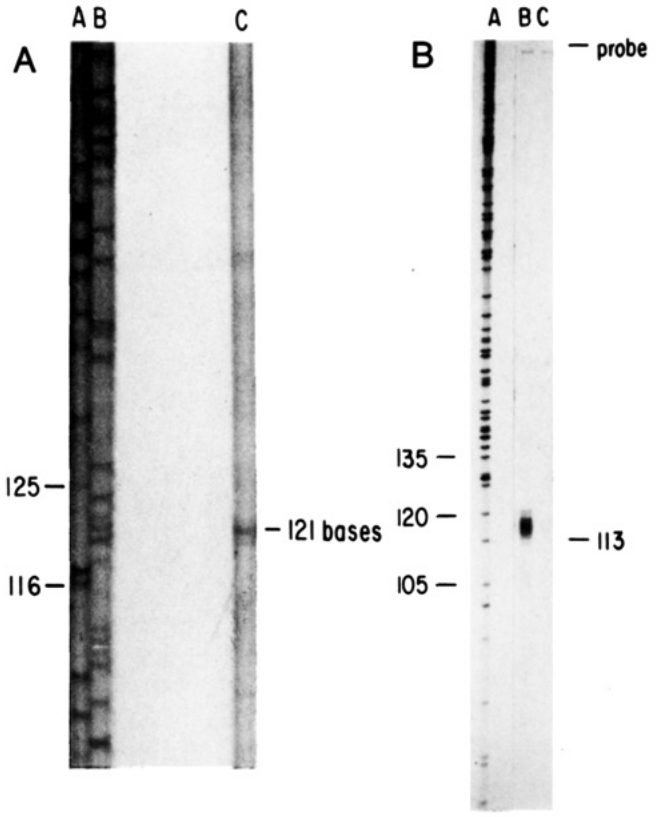

FIGURE 4: Mapping the 5′ end of the collagenase gene. (A) Lanes A and B: DNA size standards. Lane C: Primer extension product from a 20-base oligonucleotide complementary to the sequence of bases 27–46 of collagenase cDNA clone λ col-H22C. (B) Lane A: DNA size standards. Lane B: S1 nuclease protected fragment resulting from the hybridization of collagenase mRNA with an *Nco*I restriction fragment containing the 5′ end of the collagenase gene. Lane C: Results of S1 nuclease treatment of a hybridization reaction containing the denatured *Nco*I restriction fragment without RNA.

and treated with S1 nuclease. The S1-protected product was analyzed on a denaturing polyacrylamide gel. Several bands, beginning at a size of 113 bases and extending to a size of 120 bases, were observed (Figure 4B). Multiple bands are typically seen with S1 nuclease analysis possibly as a result of steric hindrance due to the presence of the cap on the 5′ end of the messenger RNA. In keeping with the primer extension experiment described above, 113 bases is the exon length required to reach the 5′ end of the collagenase message, that is, 38 bases to the end of col-H22C plus 75 bases more. Thus, these data (Figure 4) define the 5′ end of the first exon. The 113-base fragment would place the start site of collagenase gene transcription at an adenine residue typical of transcription initiation sites. Furthermore, in the predicted position, beginning 32 bases upstream of this adenine, lies a sequence, TATAAA, which conforms to the TATA homology, a DNA element thought to be important for transcription initiation (Goldberg et al., 1979).

Sequence of Collagenase cDNA and Predicted Primary Amino Acid Sequence. Sequenced here are 1926 bases from cDNA and genomic clones representing the synovial cell collagenase message (Figure 3). The message contains a single long open reading frame that codes for a protein of 468 amino acids with a predicted molecular mass of 53 679 kDa. Sixty bases of putative 5′ untranslated sequence preceed the first ATG codon which begins the open reading frame. Surrounding this methionine codon is an appropriate sequence context for efficient translation initiation (Kozak, 1986). Encoded by the first 14 amino acids at the beginning of the

Table I: Size of Exons and Introns and Sequence of Exon/Intron Junctions in the Collagenase Gene[a]

Exon/Intron	Size	5′ splice site	3′ splice site
Consensus		**AG/GTAAGT**	**(Py)$_6$NCAG/**
exon 1	164		
intron 1	about 790	AG/GTAA**A**T	CTTTT**A**TCAG/
exon 2	245		
intron 2	93	AG/GTAA**TC**	C**A**TTG**T**G**A**AG/
exon 3	149		
intron 3	235	AG/GTAAG**C**	TTTCTTG**T**AG/
exon 4	125		
intron 4	about 680	**AA**/GTAAGT	TCTCTTT**T**AG/
exon 5	156		
intron 5	160	**T**G/GTAAGT	TTTTCTT**T**AG/
exon 6	118		
intron 6	about 1800	AG/GTAAG**A**	---
exon 7	134		
intron 7	about 1900	---	---
exon 8	163		
intron 8	about 2500	AG/---	TTTT**G**ACCAG/
exon 9	104		
intron 9	175	**T**G/GTAAGT	TTTCCCTCAG/
exon 10	572		
Total Gene = 9.111 Kb			

[a] The consensus sequence for splice sites was taken from Mount (1982). Bold letters indicate deviations from this consensus.

open reading frame is a hydrophobic core structure typical of a signal peptide. From the patterns of cleavage documented for other proteins (Perlman & Halvorson, 1983), we predict the site of signal peptide cleavage is after the glycine at amino acid position 18. This would result in a mature collagenase proenzyme of 451 amino acids with a molecular mass of 51 826 kDa. Two potential sites for N-linked glycosylation (Pless & Lenarz, 1977) are located at amino acid positions 119 and 142 with the sequences Asn-Tyr-Thr and Asn-Val-Thr. Following the putative translation stop codon lie 458 bases of additional sequence representing the 3′ untranslated portion of the collagenase message. Within this region lie three repeats of a sequence, ATTTA, beginning at bases 1536, 1576, and 1621. This sequence has been previously identified in the 3′ untranslated portions of messages for lymphokines and oncogenes and has been implicated in the increased stability of mRNA after treatment of cells with phorbol esters (Shaw & Kamen, 1986). A consensus sequence for poly(A) addition begins 20 bases before the end of the untranslated region.

Homology of Rabbit Collagenase cDNA to Other cDNAs. The cDNA sequences for human skin and human synovial cell collagenase have been found to be identical (Goldberg et al., 1986; Brinckerhoff et al., 1987a). We determined the extent of homology by aligning the predicted protein sequences of rabbit and human skin collagenases (data not shown). This analysis showed 86% homology between rabbit synovial cell and human skin/synovial cell collagenase, demonstrating a very strong interspecies conservation of this enzyme.

Since rabbit proactivator/stromelysin is another metalloproteinase that plays an important role in the degradation of extracellular matrix (Vater et al., 1983; Chin et al., 1985) by (1) its ability to activate latent collagenase and (2) its ability to degrade noncollagenous matrix, we compared the cDNA-predicted amino acid sequence of rabbit collagenase to that deduced from the sequence of rabbit activator cDNA (Fini et al., 1987). Figure 5 shows the alignment of the two proteins. Amino acid conservation is 51%, suggesting that these metalloproteinases share comembership in a gene family.

Structure of the Rabbit Collagenase Gene. The exon–intron structure of the rabbit synovial cell collagenase gene, allele A, was mapped by a combination of DNA sequencing and S1 nuclease analysis (Figures 1, 6, and 7; Table I). To locate the position and lengths of exons 1–6 and 9–10, the sequenced

```
                        8          18          28    33
Activator           MKTLPTLL LLCVALCSAY PLDGASRDAD TTNMD
                     **  *     ***** ** ****  **** *****
Collagenase         MPGLPLLL LLLWGVGSHG FPA-ASETQE QDVE-
                        8          18          27    31

          43          53          63          73          83
LLQQYLENYY NLEKDVKQFV KRKDSSPVVK KIQEMQKFLG LEVTGKLDSN
** *        ** *****  ******* *  ***  * *   *    * **
MVQKYLENYY NLKDDWRKIP KQRGNGLAVE KLKQMQEFFG LKVTGKPDAE
          41          51          61          71          81

          93         103         113         123         133
TLEVIRKPRC GVPDVGHFST FPGTPKWTKT HLTYRIVNYT PDLPRDAVDA
  *****         ** ** *  * * **        *         * **  *
TLKMMKQPRC GVPDVAQFVL TPGNPRWEQT HLTYRIENYT PDLSRADVDN
          91         101         111         121         131

         143         153         163         173         183
AIEKALKVWE EVTPLTFSRK YEGEADIMIS FGVREHGDFI PFDGPGNVLA
    *** * *      *** ** *       **** * **      ***
AIEKAFQLWS NVTPLTFTKV SKGQADIMIS FVRGDHRDNS PFDGPEGQLA
         141         151         161         171         181

         193         203         213         223         233
HAYAPGPGIN GDAHFDDDEQ WTKDTTGTNL FLVAAHELGH SLGLFHSANP
  **  *  *   *  * **     ****   **              *  ***
HAFQPGLGIG GDVHFDEDDR WTKDFRNYNL YRVAAHELGH SLGLSHSTDI
         191         201         211         221         231

         243         253         263         273         283
EALMYPVYNA FTDLARFRLS QDDVDGIQSL YGPAPASPDN SGVPMEPVPP
*    **** ******** *    *    **    *******  ** **  **
GALMYPN--- YMFSGDVQLA QDDIDGIQAI YGP------- SQNPSQPV--
    238         248         258                 269

         293         303         313         323         333
GSGTPVMCDP DLSFDAISTL RGEILFFKDR YFWRKSLRIL EPEFHLISSF
 **  **  * * *   * *     *       *** ******  * ***  *
GPQTPKVCDS KLTFDAITTI RGEIMFFKDR FYMRANPYYS EVELNFISVF
         279         289         299         309         319

         343         353         363         373         382
WPSLPSAVDA AYEVISRDTV FIFKGTQFWA IRGNEVQAGY PRSIH-TLGF
  *  ****      **  ** **   *** * ** *****    **  ***
WPHLPNGLQA AYEVAHRDEI LFFKGNKYWT VQGQNELPGY PKDIHSSFGF
         329         339         349         359         369

         392         402         412         422         432
PSTIRKIDAA ISDKERKKTY FFVEDKYWRF DEKRQSLEPG FPRHIAEDFP
 *****     * *****       **     *   *** ***  * ** **
PRSVNHIDAA VSEEDTGKTY FFVANKYWRY DEYKRSMDAG YPKMIEYDFP
         379         389         399         409         419

         442         452         462         472
GINPKIDAVF EAFGFFYFFS GSSQSEFDPN AKKVTHVLKS NSWFQC      478
  **  *     ***       *  ** **   * * ****** *     * ***
GIGNKVDAVF KKDGFFYFFH GTRQYKFDPK TKRILTLQKA NSWFNCRKN   468
         429         439         449         459
```

FIGURE 5: Amino acid sequence alignment of activator and collagenase. The deduced amino acid sequence of rabbit preprocollagenase is aligned with that of rabbit preproactivator. Stars indicate amino acids that do not match between the two proteins.

region was examined for colinearity with the allele A cDNA. No sequence differences were found between the cDNA and genomic sequence. Intron/exon boundries were defined by the termination of colinearity with the cDNA, and where some overlap of sequence was found between two exons, by comparison to the published consensus for the splice junction (Mount, 1982) (Table I and Figure 1). The lengths of introns 2, 3, 5, and 9 were taken as the length of the DNA sequence between each exon. In the case of intron 1, the length was deduced by mapping the position on the genomic clones of an *Nco*I site found within the first exon (Figure 1). Similarly, the length of intron 4 was determined by mapping the position of a *Bst*XI site found within exon 4 (Figure 1).

The preceding analysis mapped exon positions within the gene for all but 300 bases toward the 3′ end of the cDNA. Preliminary mRNA hybridization experiments of mRNA from phorbol ester induced synovial cells to fragments of genomic clones (Fini et al., 1986b) indicated that at least two more exons were required to encode this portion of the message, one within the 1.5-kb *Eco*RI fragment and one within the 3.05-kb *Eco*RI fragment of the genomic clones (see Figure 1). To determine the length and number of the remaining exons

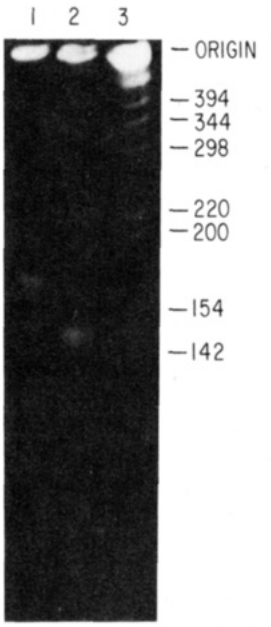

FIGURE 6: S1 mapping of collagenase gene exons 7 and 8. S1 nuclease digestion products of collagenase mRNA with single-stranded M13 subclones containing collagenase gene fragments. As explained in the text, the size of the protected DNA–RNA hybrid defines the size of the collagenase gene exon found in the examined region of the gene. Lane 1: Size of exon 8. Lane 2: Size of exon 7. Lane 3: Size standards.

contained within these fragments, S1 analysis was performed (Figure 6). Single-stranded subclones of the *Eco*RI fragments of 1.5 and 3.05 kb in M13 were annealed to the opposite strand of the cDNA, also cloned in M13, and treated with S1 nuclease to digest the unprotected regions. A cDNA fragment of approximately 135 bases was protected by the *Eco*RI fragment of 1.5 kb, and a fragment of about 160 bases was protected by the *Eco*RI fragment of 3.05 kb. This defined the lengths of exons 7 and 8 and accounted for the final 300 bases in the cDNA. We have examined the cDNA sequence around the approximate position of splicing of exons 7 and 8. To most closely conform to the sequence requirements identified for RNA splicing (Mount, 1982), exon 7 should be 134 bases and exon 8 should be 163 bases in length, and they are identified as such in Table I. Also, the splicing sequence considerations were followed in delineating the position of the junction between exon 6/7 and 7/8 in Figure 3.

The lengths of introns 6–8 were determined by further restriction mapping of the genomic clones to locate the position of exons 7 and 8. A single site for the restriction enzyme *Sst*I is found in the *Eco*RI fragment of 1.5 kb. Since there is also a site for *Sst*I in exon 7, this localizes the position of the exon within the *Eco*RI 1.5-kb fragment. *Sca*I, which cleaves at the beginning and end of exon 8, was found to cleave the 3.05-kb *Eco*RI genomic fragment also in only two places.

The sequence of the new cDNA clones described here shows no evidence of a poly(A) tail. However, the cDNA clone, H9, which contains an additional eight bases of cDNA at the 3′ end, is terminated with a run of four adenine residues (Fini et al., 1986b). Furthermore, the presence of a poly(A) addition signal beginning 24 bases before the adenine run (20 bases before the end of the new cDNA clones) also suggests that we have succeeded in cloning very close to the 3′ end of the message. The size of the collagenase message on agarose gels, 2.2 kb, could be predicted from the size of the sequenced message added to a poly(A) tail of 150–200 bases. We have described experiments showing that exon 10 is not interrupted by intervening sequences and defining the end of the col-

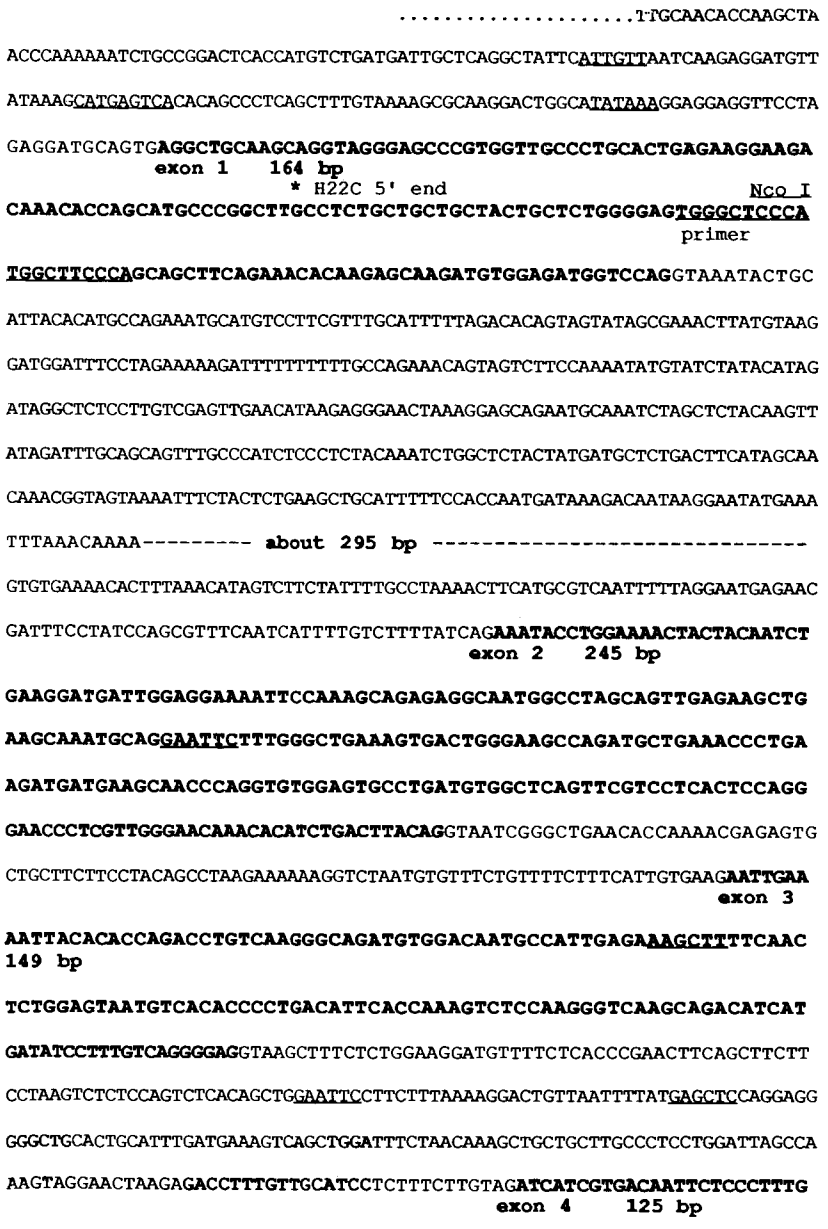

FIGURE 7: Partial sequence of the synovial cell collagenase gene. Exons are indicated in bold letters. Underlined are those restriction sites found on the detailed map of λ 13A in Figure 1, except the *Nco*I site used for S1 mapping in Figure 4B, which is overlined. The sequence complementary to the oligonucleotide used for primer extension in Figure 4A is delineated. The TATA box beginning 32 bases before the first base of exon 1 is underlined. Also underlined are the CAT box at position −105 to −100 and, at position −78 to −70, the sequence homologous to the 5′ flanking DNA of the rat transin gene.

lagenase gene within a 1.5-kb region on the genome that includes exon 10 (Fini et al., 1986b). Because of these data, we have given the length of exon 10 in Table I as the distance to the base before the run of four adenines in H9.

Discussion

The synovial cell collagenase described here degrades the interstitial collagens types I–III and has similar properties to the interstitial collagenase of skin and of macrophages (Harris et al., 1984). In fact, recent data argue strongly that the skin and synovial cell collagenases of human are products of the same gene (Goldberg et al., 1986; Brinckerhoff et al., 1987a). However, other collagenases, in particular, the enzyme from neutrophils (Hasty et al., 1984) and uterus (Welgus et al., 1985) and the type IV collagen degrading enzyme from tumor cells (Salo et al., 1983), differ in molecular weight, isoelectric point, immunological identity, and substrate preference. These enzymes may well prove to be products of their own separate genes. Sequence analysis of rabbit synovial cell collagenase and human skin/synovial cell collagenase indicates that they share approximately 86% homology, while rabbit collagenase and activator/stromelysin share about 50% homology. This nucleic acid sequence homology, together with the requirement of these enzymes for metal ions (Barrett, 1977), allows us to assign rabbit synovial cell collagenase, along with human skin/synovial cell collagenase and rabbit synovial cell activator/stromelysin to a gene family of metalloproteinases that degrade the connective tissue matrix (Fini et al., 1987). On the basis of their nucleic acid homology, human stromelysin (Whitham et al., 1986) and rat transin, an oncogene-induced protein (Matrisian et al., 1985), are also members of this family.

Indeed, transin and stromelysin are strongly homologous and are probably interspecies homologues (Matrisian et al., 1985; Whitham et al., 1986; Fini et al., 1987). Examination of the sequences of collagenase in rabbit and human (Goldberg et al., 1986), and of rabbit activator (Fini et al., 1987) and its homologues (Matrisian et al., 1985; Whitham et al., 1986), reveals certain structural features that are conserved among the different proteins, suggesting their functional importance and demonstrating the close relationship of these metalloproteinases. These structural similarities include strong sequence homology, sites for N-linked glycosylation, and the conservation of three cysteine residues of potential importance in tertiary structure. While these features are strongly conserved in collagenase, activator, transin, and stromelysin, they are present to a more limited extent in the heme-binding protein homopexin (Takahashi et al., 1985; Matrisian et al., 1985), suggesting a more distant relationship of the gene for this protein to the others. By contrast, the metalloproteinase carboxypeptidase A (Quinto et al., 1982) shows none of these homologies, demonstrating its membership in a separate enzyme subclass.

In the rabbit, at least two collagenase genes of the synovial cell type are found at two different genetic loci (Fini et al., 1986b). Here, we have determined the structure of one of these genes. Its entire length is 9.1 kb, and there are a total of 10 exons and 9 introns. The length of the introns varies between 0.1 and 2.5 kb. Exons 1 and 3–9 are very similar in size and have an average length of 140 base pairs, which is very close to the average value reported for eucaryotic exons (Blake, 1983). Exon 2, at 244 bases, is somewhat larger than the other exons, but the largest exon is the 10th with a length of 572 bases. The unusually long length of the final exon is characteristic of many eucaryotic genes including the serine proteinase genes of mammals (Rogers, 1985).

The exon/intron structure of rat transin genes I and II has been recently determined (Breathnach et al., 1987). These genes show remarkable similarity in structure to the rabbit synovial cell collagenase gene. All three genes are split by nine introns, and all the introns are inserted in the portion of the gene that encodes protein. The position within the amino acid coding regions where introns are inserted are in identical places in both transin genes and also in identical places in the rabbit collagenase gene—a gene for a proteinase with a different substrate specificity than rat transin and found in a different species. Conservation of intron location within the coding portion of individual members of a multigene family is a striking and frequently found phenomenon (Estratiadis et al., 1980). The structural similarity between these genes substantiates the hypothesis that they evolved from a common ancestor which possessed this exon/intron relationship.

Analysis of the nucleic acid sequence of the exons allows us to deduce several structural/functional features of the collagenase protein. Exon 1 codes for the 5′ untranslated portion of the messenger RNA as well as for the signal peptide. It terminates 16 codons following the proposed site for signal cleavage. Exon 2 probably contains the site of proenzyme cleavage following activation. The N-terminus of the activated human enzyme has been determined by protein sequencing (Goldberg et al., 1986; Whitham et al., 1986). It lies 99–100 amino acids from the cDNA-deduced N-terminus of the human preproenzyme. By analogy, the cleavage site activating the highly homologous rabbit enzyme must lie very close to this position. Exon 3 contains both of the potential sites for N-linked glycosylation found in the protein. It is not known whether both are used. Exon 5 contains a potential site for binding of zinc, an ion necessary for collagenase function, as recognized by homology to bacterial enzymes (Whitham et al., 1986). No function can as yet be assigned for exon 4 or 6–9 although they may prove to have functional identities once the amino acids required for enzyme activity, collagen binding, and Ca^{2+} ion binding are determined. Of interest, however, is the fact that the amino acid sequence encoded by exons 6–10 has a weak homology to the heme-binding protein homopexin (Takahashi et al., 1985). It is tempting to speculate that this portion of the collagenase gene has a similar binding function. It could be involved in recognition of the enzyme substrate, collagen, in binding to inhibitors of enzyme activity, or in anchoring the enzyme to the connective tissue matrix at its site of action. Exon 10 codes for the last 36 amino acids of collagenase followed by the termination codon. The remainder of this long exon is the 3′ untranslated sequence.

One of our primary aims for determining the sequence of the collagenase gene and message was the opportunity to examine this sequence for DNA elements reported to be important in regulation of gene expression. Of interest therefore are the three repetitions of a sequence, ATTTA, found in the 3′ untranslated portion of the rabbit collagenase message. This sequence has been found in the 3′ untranslated region of messages for certain inflammatory proteins and oncogenes (Shaw & Kamen, 1986). Recent experiments have implicated the sequence in control of the rapid degradation of mRNA to which it is linked and in the stabilization of this mRNA in cells treated with phorbol esters. Of importance is the conservation of all three repetitions of the ATTTA box in the 3′ untranslated region of human collagenase mRNA. The element is also repeated once in the coregulated rabbit activator/stromelysin mRNA and twice in the rat transin message. Our recently published observation that collagenase mRNA is stabilized in cells treated with phorbol esters suggests a

potential function for the ATTTA box in regulation of collagenase and activator message levels (Brinckerhoff et al., 1986).

We have described above the sequencing of 172 bases of DNA flanking the start site of collagenase gene transcription. We have examined this sequence for DNA elements reported in the literature to have functional importance in the initiation and efficiency of transcription. Beginning thirty-two bases upstream of the start site for transcription lies a sequence, 5′-TATAAA-3′, homologous to the TATA box implicated in accuracy of transcription initiation. In addition, a six-base element, 5′-ATTGTT-3′, is found at bases −105 to −100. This sequence, read 5′ to 3′ from the commplementary DNA strand, is AACAAT, a CAT box (Bucher & Trifonov, 1986), an element shown to be important for transcription of globin genes. A comparison of the 5′ flanking DNA sequence of the rat transin gene (Breathnach, 1987) with that of the rabbit collagenase gene should be particularly revealing since these genes are regulated by similar inducing agents including the tumor-promoting phorbol esters. Interestingly, a sequence, 5′-CATGAGTCA-3′, located at bases −78 to −70 in the rabbit, is conserved in the rat transin gene at positions −72 to −65, and in the rat transin II at positions −59 to −51. This finding strongly suggests that the sequence may function as a binding site for transcriptional regulatory proteins, much like that for the glucocorticoids (Eliard et al., 1985). This element has some homology to a DNA sequence that was functionally identified as a phorbol ester response element (Comb et al., 1986). Functional assays will be essential to determine both the true identity of this homology as a transcriptional response element and the nature of the inducers that may interact with it.

Added in Proof

While this paper was in press, Angel et al. (1987) reported the same sequence motif, i.e., 5′-CATGAGTCAG-3′, around region −70 of the 5′ end of the human collagenase gene, and demonstrate the importance of this sequence in the regulation of gene expression by phorbol esters. Furthermore, a comparison of the DNA flanking the rabbit collagenase gene with that flanking the human collagenase gene shows a very high degree of sequence conservation, implicating additional portions of this DNA in gene regulation.

Acknowledgments

We thank Elizabeth Lombardi, Dr. John Vournakis, and the Molecular Genetics Center of Dartmouth College for preparation of synthetic oligonucleotides used in this work. We also thank Dr. Ross C. Hardison of the Pennsylvania State University for his gift of the rabbit genomic library.

References

Angel et al. (1987) *Cell (Cambridge, Mass.) 49*, 729–739.

Barrett, A. J. (1977) *Proteinases in Mammalian Cells and Tissues* (Barrett, A. J., Ed.) North-Holland, New York.

Blake, C. (1983) *Nature (London) 306*, 535–537.

Breathnach, R. Matrisian, L. M., Gesnel, M. C., Staub, A., & Leroy, P. (1987) *Nucleic Acids Res. 15*, 1139–1151.

Brinckerhoff, C. E., & Harris, E. D., Jr. (1981) *Biochim. Biophys. Acta 677*, 424–432.

Brinckerhoff, C. E., McMillan, R. M., Fahey, J. V., & Harris, E. D., Jr. (1979) *Arthritis Rheum. 22*, 1109–1116.

Brinckerhoff, C. E., McMillan, R. M., Dayer, J.-M., & Harris, E. D., Jr. (1980) *N. Engl. J. Med. 303*, 432–435.

Brinckerhoff, C. E., Nagase, H., Nagle, J. E., & Harris, E. D., Jr. (1981) *J. Am. Acad. Dermatol. 6*, 591–602.

Brinckerhoff, C. E., Gross, R. H., Nagase, H., Sheldon, L. A., Jackson, R. C., & Harris, E. D., Jr. (1982) *Biochemistry 21*, 2674–2679.

Brinckerhoff, C. E., Plucinska, I. M., Sheldon, L. A., & O'-Connor, G. T. (1986) *Biochemistry 25*, 6378–6384.

Brinckerhoff, C. E., Ruby, P. L., Austin, S. D., Fini, M. E., & White, H. D. (1987a) *J. Clin. Invest. 79*, 542–546.

Brinckerhoff, C. E., Fini, M. E., Ruby, P. L., Plucinska, I. M., Borges, K. A., & Karmilowicz, M. J. (1987b) in *Development and Diseases of Cartilage and Bone Matrix*, pp 299–317, Liss, New York.

Bucher, P., & Trifonov, E. N. (1986) *Nucleic Acids Res. 14*, 10009–10027.

Chin, J. R., Murphy, G., & Werb, Z. (1985) *J. Biol. Chem. 260*, 12367–12376.

Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., & Rutter, W. J. (1979) *Biochemistry 18*, 5294–5299.

Comb, M., Bernbirg, N. C., Seasholtz, A., Herbert, E., & Goodman, H. M. (1986) *Nature (London) 323*, 353–356.

Dale, R. M. K., McClure, B. A., & Houchins, J. P. (1985) *Plasmid 13*, 31–40.

Dayer, J.-M., Krane, S. M., Russell, R. G. G., & Robinson, D. R. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 945–949.

Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connell, C., Spritz, R. A., DyldeRiel, J. K., Forget, D. G., Weissman, S. M., Slighton, J. L., Smithies, O., Barelle, F. E., Shoulders, C. C., & Proudfoot, N. J. (1980) *Cell (Cambridge, Mass.) 21*, 653–668.

Eliard, P. H., Marchand, M. J., Rousseau, G. G., Formstecher, P., Mathy-Hartert, M., Belayew, A., & Martial, J. A. (1985) *DNA 4*, 409–417.

Favoloro, J., Treisman, R., & Kamen, R. (1980) *Methods Enzymol. 65*, 718–749.

Fini, M. E., Austin, S. A., Holt, P. T., Ruby, P. L., Gross, R. H., White, H. D., & Brinckerhoff, C. E. (1986a) *Collagen Relat. Res. 6*, 239–248.

Fini, M. E., Gross, R. H., & Brinckerhoff, C. E. (1986b) *Arthritis Rheum. 29*, 1301–1315.

Fini, M. E., Karmilowicz, M. J., Ruby, P. L., Beeman, A. M., Borges, K. A., & Brinckerhoff, C. E. (1987) *Arthritis Rheum.* (in press).

Goldberg, G. I., Wilhelm, S. M., Kronberger, A., Bauer, E. A., Grant, G. A., & Eisen, A. Z. (1986) *J. Biol. Chem. 261*, 6600–6605.

Goldberg, M. (1979) Dissertation, Stanford University, Stanford, CA.

Gross, J. (1982) in *Cell Biology of the Extracellular Matrix* (Hay, E., Ed.) pp 217–253, Plenum, New York.

Gross, R. H. (1986) *Nucleic Acids Res. 14*, 591–596.

Gross, R. H., Sheldon, L. A., Fletcher, C. F., & Brinckerhoff, C. E. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1981–1985.

Harris, E. D., Jr. (1985) in *Textbook of Rheumatology* (Kelley, W. N., Harris, E. D., Jr., Ruddy, S., & Sledge, C. B., Eds.) pp 886–914, W. B. Saunders, Philadelphia.

Harris, E. D., Jr., DiBona, D. R., & Krane, S. M. (1969) *J. Clin. Invest. 48*, 2104–2113.

Hasty, K. A., Hibbs, M. S., Kang, A. H., & Mainardi, C. L. (1984) *J. Exp. Med. 159*, 1455–1463.

Hay, E. D. (1984) in *Cell Biology of the Extracellular Matrix* (Hay, E., Ed.) pp 1–32, Plenum, New York.

Kozak, M. (1986) *Cell (Cambridge, Mass.) 44*, 283–292.

Liotta, L. A., Rao, C. N., & Barsky, S. H. (1984) in *The Role of Extracellular Matrix in Development* (Trelstad, R. L., Ed.) pp 357–372, Liss, New York.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual*, Cold Spring

Harbor Laboratory, Cold Spring Harbor, NY.
Matrisian, L. M., Glaichenhaus, N., Gesnel, M.-C., & Breathnach, P. (1985) *EMBO J. 4*, 1435–1440.
Matrisian, L. M., LeRoy, P., Ruhlmann, C., Gesnel, M. C., & Breathnach, R. (1986) *Mol. Cell. Biol. 6*, 1679–1686.
Mignatti, P., Robbins, E., & Rifkin, D. B. (1986) *Cell (Cambridge, Mass.) 47*, 487–498.
Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
Nagase, H., Brinckerhoff, C. E., Vater, C. A., & Harris, E. D., Jr. (1983) *Biochem. J. 214*, 281–288.
Perlman, D., & Halvorson, H. O. (1983) *J. Mol. Biol. 167*, 391–409.
Pless, D. D., & Lenarz, W. J. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 134–138.
Quinto, C., Quiroga, M., Swain, W. F., Nikovits, W. C., Jr., Standring, D. N., Pictet, R. L., Valenzuela, P., & Rutter, W. J. (1982) *Proc. Natl. Acad. Sci. US.A. 79*, 31–35.
Rogers, J. (1985) *Nature (London) 315*, 458–459.
Salo, T., Liotta, T., & Tryggvason, K. (1983) *J. Biol. Chem. 258*, 3058–3063.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*; 5463–5467.
Shaw, G., & Kamen, R. (1986) *Cell (Cambridge, Mass.) 46*, 659–667.
Takahashi, N., Takahashi, Y., & Putnam, F. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 73–77.
Ullrich, A., Shine, J., Chirgwin, J., Pictet, W. J., Tischer, E., Rutter, W. J., & Goodman, H. M. (1977) *Science (Washington, D.C.) 196*, 1313–1319.
Vater, C. A., Nagase, H., & Harris, E. D., Jr. (1983) *J. Biol. Chem. 258*, 9374–9382.
Welgus, H. G., Gant, G. A., Sacchettini, J. C., Roswit, W. T., & Jeffrey, J. J. (1985) *J. Biol. Chem. 260*, 13601–13606.
Whitham, S. E., Murphy, G., Angel, P., Rahmsdorf, H.-J., Smith, B. J., Lyons, A., Harris, T. J. R., Reynolds, J. J., Herrlich, P., & Docherty, A. J. P. (1986) *Biochem. J. 240*, 913–916.
Wooley, D. E., & Evanson, J. J. (1980) *Collagenase in Normal and Pathological Connective Tissues*, Wiley, New York.

Nucleotide Sequence of the Gene for Human Prothrombin†

Sandra J. Friezner Degen*,‡ and Earl W. Davie§

Children's Hospital Research Foundation and Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio 45229, and Department of Biochemistry, University of Washington, Seattle, Washington 98195

Received April 16, 1987; Revised Manuscript Received May 29, 1987

ABSTRACT: A human genomic DNA library was screened for the gene coding for human prothrombin with a cDNA coding for the human protein. Eighty-one positive λ phage were identified, and three were chosen for further characterization. These three phage hybridized with 5′ and/or 3′ probes prepared from the prothrombin cDNA. The complete DNA sequence of 21 kilobases of the human prothrombin gene was determined and included a 4.9-kilobase region that was previously sequenced. The gene for human prothrombin contains 14 exons separated by 13 intervening sequences. The exons range in size from 25 to 315 base pairs, while the introns range from 84 to 9447 base pairs. Ninety percent of the gene is composed of intervening sequence. All the intron splice junctions are consistent with sequences found in other eukaryotic genes, except for the presence of GC rather than GT on the 5′ end of intervening sequence L. Thirty copies of *Alu* repetitive DNA and two copies of partial *Kpn*I repeats were identified in clusters within several of the intervening sequences, and these repeats represent 40% of the DNA sequence of the gene. The size, distribution, and sequence homology of the introns within the gene were then compared to those of the genes for the other vitamin K dependent proteins and several other serine proteases.

The coagulation of blood in mammals is the result of many enzymatic reactions involving the sequential activation of a number of specific serine proteases by limited proteolysis. The end result of this series of reactions is the formation of an insoluble fibrin clot (Davie et al., 1979). Prothrombin participates in the final stage of this process when it is activated to thrombin by factor Xa in the presence of factor Va, calcium ions, and a phospholipid surface. Thrombin, in turn, cleaves fibrinopeptides A and B from the α and β chains of fibrinogen, respectively, to form the fibrin clot. Thrombin is also involved in the activation of factors V, VIII, and XIII and protein C, in the stimulation of platelets to undergo a change of shape, and in the regulation of proliferation of certain cell types, most notably endothelial cells (Davie et al., 1979; Esmon, 1983; Davey & Luscher, 1967; Chen & Buchanan, 1975).

Prothrombin is synthesized in the liver as a single-chain glycoprotein with a M_r of 72 000 (Barnhart, 1960; Mann & Elion, 1980). Vitamin K is required for the γ-carboxylation of prothrombin, as well as several other plasma proteins participating in coagulation. The carboxylation involves 10 amino-terminal glutamic acid residues that are converted to γ-carboxyglutamic acid (Gla;[1] Stenflo et al., 1974; Nelsestuen et al., 1974; Magnusson et al., 1975). These residues are required for prothrombin to bind calcium and thus aid in the binding of the protein to phospholipid surfaces. The biosyn-

†This work was supported in part by the Pew Memorial Trust, by Research Grants HL16919 and HL38232 from the National Institutes of Health, and by start-up funds from the Children's Hospital Research Foundation, Cincinnati, OH. S.J.F.D. is a Pew Scholar in the Biomedical Sciences.

*Author to whom correspondence should be addressed.
‡University of Cincinnati College of Medicine.
§University of Washington.

[1] Abbreviations: bp, base pairs; kb, kilobase pairs; Gla, γ-carboxyglutamic acid; t-PA, tissue plasminogen activator; u-PA, urokinase-type plasminogen activator; EDTA, ethylenediaminetetraacetic acid; SDS, sodium dodecyl sulfate.

0006-2960/87/0426-6165$01.50/0